INFLATION, EMPLOYMENT AND BUSINESS FLUCTUATIONS

Volume C

Edmond Malinvaud

MACROECONOMIC THEORY

A Textbook on Macroeconomic Knowledge and Analysis

VOLUME C: INFLATION, EMPLOYMENT AND BUSINESS FLUCTUATIONS

ADVANCED TEXTBOOKS IN ECONOMICS

VOLUME 35

ELSEVIER

Amsterdam–London–New York–Oxford–Paris–Shannon–Tokyo

MACROECONOMIC THEORY
A Textbook on Macroeconomic Knowledge and Analysis

Edmond MALINVAUD

Volume C: INFLATION, EMPLOYMENT AND BUSINESS FLUCTUATIONS

Adaptation and revision of *Théorie Macroéconomique*, published in 1981–1982 by Dunod, Paris, and translated into English by Fenella Kirman.

N·H

2000

ELSEVIER
Amsterdam–London–New York–Oxford–Paris–Shannon–Tokyo

ELSEVIER SCIENCE B.V.
Sara Burgerhartstraat 25
P.O. Box 211, 1000 AE Amsterdam, The Netherlands

First edition 2000

Library of Congress Cataloging in Publication Data
A catalog record from the Library of Congress has been applied for.

ISBN: 0-444-82862-1 (Volume 35A); 0-444-82863-X (Volume 35B); 0-444-50518-0 (Volume 35C); 0-444-88497-1 (set)

ISSN: 0169-5568 (Series)

©The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

INTRODUCTION TO THE SERIES

The aim of the series is to cover topics in economics, mathematical economics and econometrics, at a level suitable for graduate students or final year undergraduates specializing in economics. There is at any time much material that has become well established in journal papers and discussion series which still awaits a clear, self-contained treatment that can easily be mastered by students without considerable preparation or extra reading. Leading specialists will be invited to contribute volumes to fill such gaps. Primary emphasis will be placed on clarity, comprehensive coverage of sensibly defined areas, and insight into fundamentals, but original ideas will be not be excluded. Certain volumes will therefore add to existing knowledge, while others will serve as a means of communicating both known and new ideas in a way that will inspire and attract students not already familiar with the subject matter concerned.

The Editors

This Page Intentionally Left Blank

Preface

This textbook may appear uncommon because of its length and its organization, because of its eclectic blend of approaches, empirical claims and modelling developments, finally because of its frequent reference to old material and indeed because of the age of its author. Readers will judge from the results. Here are, however, a few indications which may explain part of the singularity and suggest how to best use the book.

It is the outcome of the last forty years during which the author worked on macroeconomics, teaching it, building tools for its study, particularly within the public service, or advising on policies. This activity took place in various environments; it always led to reflect on the content of macroeconomic knowledge, as well as on research methods for improving or applying knowledge; it called for continually reading the literature, often marvelling at new ideas or research innovations, sometimes also annoyed at misplaced claims or risky fashions, always wondering how recent results were progressively enriching or altering earlier accumulated knowledge.

Such a long maturation inevitably entails a perspective in which contributions made in the 1950s and 1960s find their place beside more modern ones, even when the latter may claim to encompass some of the former. But the book is not about the various macroeconomic schools, nor about the history of macroeconomic ideas. It rather aims at offering an organized treatment of present knowledge, occasionally exposing the contributions of famous authors or famous works, but only when that seemed to be unavoidable.

If the balance between the places respectively given to various topics or approaches differs from that found in other books, it may also be because of a different perception by the author on what we really know, on how we could know more, on what would be crucial for a fuller and more relevant knowledge. So are involved ideas about the purpose of knowledge and ideas about the methods that may make it more reliable.

This book may appear long on preliminaries. Indeed, questions concerning economic growth and business fluctuations will not be really tackled before Chapter 5, that is after a good third of the whole text. The reason will be presented in the introduction. In any case such a long book is bound to be often used like an encyclopedia. The student who would decide to begin reading at Chapter 5 would not be lost, if he or she already had a good background. The student or teacher who would have once understood how arguments and evidences were developing throughout the full book could usefully refer back to particular sections for closer examination.

The book is addressed mainly to students who are already acquainted with economics and its methods. No prerequisite is absolutely required; but the text is too dense for being directly accessible to those who would be ignorant of economic concepts, of economic phenomena or of the economics discipline. The author hopes that the book will be used in the first years of post-graduate teaching and that it will serve to all kinds of economists as a reference on particular macroeconomic issues or approaches with which those economists may not be familiar. No one will be surprised if he or she finds in the book a slight bias in favour of questions or models at which the author worked.

The book is organized in nine long chapters, divided into parts and further into sections. It is made of three volumes. Volume A (Chapters 1–4) presents the conceptual framework of macroeconomic theory and an extensive treatment of the behaviour of the two main kinds of agents, consumers and producers, seen from the macroeconomic viewpoint. Volume B (Chapters 5–7) deals with economic growth and with the static analysis of short-run changes in employment and other macroeconomic variables. Volume C (Chapters 8 and 9) is devoted to the dynamics of inflation, to business fluctuations and to macroeconomic policies.

CONTENTS OF VOLUMES A, B AND C

VOLUME A: FRAMEWORK, HOUSEHOLDS AND FIRMS

VOLUME B: ECONOMIC GROWTH AND SHORT-TERM EQUILIBRIUM

VOLUME C: INFLATION, EMPLOYMENT AND BUSINESS FLUCTUATIONS

Introduction

For many of the questions they are dealing with, economists argue within conceptual representations that do not claim to individually identify each agent and each operation. The argument can then be said macroeconomic. Some phenomena are also said to be macroeconomic because they concern the whole economy.

This book aims at teaching the knowledge of such phenomena, or rather at teaching the methods, empirical observations, questions and models used for knowledge-building about the phenomena. The domain is definitely less well unified than that covered by the microeconomic theory of prices and allocation of resources. Teaching requires an organization for the exploration of the field and this will here take the reader through a long route. A few indications about the journey have to be given from the start.

1. Phenomena to be studied

From one decade to the next neither the volume of aggregate output nor that of output per capita grow at a constant and universal pace: growth was faster at some times than at others, in some countries than in others. Long periods of approximate stagnation, even in some cases regress, have been observed. The same rate of output growth may not imply the same performance in terms of welfare, of social development or of future prospects.

Economic evolution is also more or less steady. Fairly constant and regular trends occur at times and places, from one year to the next; important fluctuations with booms and depressions are observed at other times or places. Variations may then concern the speed of change of the overall price level, or of real wage rates, as well as that of output or employment.

These two main types of phenomena, the long-run and short-run ones, are distinguished in public perception and still more systematically in economic analysis. Their empirical characterization brings out different aspects; the reasons given for them may much differ; the purposes of the explanation may be quite distinct. In particular, economic policy makers, whose demands often inspired macroeconomic research, may be contemplating different sorts of interventions, on the one hand, for improving long-run performance or, on the other hand, for curing unsatisfactory features of current evolution.

2. The role of theory

The word "theory" is understood here in its broad sense of systematic knowledge. More precisely, a theory is not meant to be simply a model, or set of models, more or less rigorously formalized. It is also a system of concepts defining the correspondence between the variables of the model and the world to which it is supposed to be applied. A theory also contains at least rough orders of magnitude for some parameters and it contains assertions about the shape of some functions. A theory finally provides a user guide explaining when and how the verbal exposition and its model(s) have to be read and applied. In other words, a theory offers a structured system of ideas and facts, which summarizes our knowledge about a phenomenon and therefore is meant to have some relevance and power in explaining causal relations in the real world.

Economic phenomena are results of human activities in societies in which they occur. In order to understand these phenomena we need to consider: (i) on the one hand, actions of participants in economic activity, i.e. mainly individual persons or households and enterprises or other similarly organized elementary decision-units, (ii) on the other hand, whichever interdependence between these actions is implied by exchanges and more generally by mutual constraints or relations in a given social and institutional context. This dual perspective is everywhere present in economic theory, within which we commonly speak of two main parts, fairly distinct from one another.

Microeconomic theory applies itself to knowing how are relative prices and remuneration rates determined and how simultaneously are productive resources allocated amongst various uses: production and distribution of goods and services, for consumption or investment. The theory pays particular attention to conditions for efficiency in this allocation of resources. It very often refers to an idealized set of institutions, called "perfect competition". It is built up by means of abstract models supposed to separately identify each good, each service, each consumer, each producer…

Macroeconomic theory intends to find out how are determined the global characteristics of economic growth and business fluctuations: volumes of the main aggregates (production, consumption, investment), rate of unemployment, functional shares in the distribution of incomes, rate of inflation... It often focuses attention on particular phenomena which are felt to play a strategic role in the determination: capital accumulation, income distribution, formation of aggregate demand, financing of operations... It claims to refer in its arguments to the main features of actual institutional contexts. It is content with models in which agents are not individually identified, the effects of their actions being captured through "*behavioural laws*", expressed at more or less high levels of aggregation, for large groups of households or firms.

However, distance between the two main parts of economic theory must not be overrated. In order to well define the nature and significance of their behavioural laws, in order to endow these laws with a few definite properties, macroeconomists have to proceed to preliminary microeconomic analyses and to study how aggregation of individual actions and behaviours operates. They have also to carefully consider the microeconomics of the institutional context they claim to represent in their models. In other words, macroeconomic theory has microeconomic foundations.

Conversely, good models of microeconomic theory, however penetrating for conceptual understanding, are suited to few applications beyond the levels of a few individual agents or a few elementary markets. They contain too many unknown parameters; their general comparative properties are too little determinate. For assessments about most economic effects of interest, we therefore need somewhat aggregated representations.

3. From observation to models

Scientific knowledge comes from observation, even though most past observations were often so much processed within models that we may lose sight of the many origins in the real world from which a particular theory has been derived. Economics is no exception. For macroeconomics, two important sources of information exist.

Many economists, from far in the past to the present, have been or are taking the option of building macroeconomic knowledge only from macroeconomic observation, as it appears in the statistics or national accounts. But there are two difficulties with such a direct inductive approach. In the first place, macroeconomic data do not come from controlled experiments but from observation of actual evolutions as they occur. One may suspect these evolutions result from many causes, often interfering with one another. The economist or econometrician who wants to learn from the data has a difficult time in identifying all the main factors, so as not to be misled by the spurious effects coming from omitted factors. In the second place, the database provided by macroeconomic observations is fairly poor, particularly so when considered against the complexity and possible lack of permanence of economic phenomena occurring in societies where institutions evolve and behaviour may change. Any rough comparison with natural sciences shows how handicapped economists are because of these two difficulties.

Fortunately, economists have an advantage over their colleagues specializing in those sciences: they have direct knowledge of economic agents and of the institutional framework in which they operate. The nature of the constraints restricting actions of firms and households is known and can be specified in the situations considered by a particular theory. The same applies to the objectives agents are trying to achieve. It is thus required for scientific efficiency that, when macroeconomic data are analyzed, our direct knowledge of institutions and behaviours is also taken into account. This is naturally done by modelling at the microeconomic level.

The art of macroeconomics is indeed to take advantage of all relevant observations and to best combine their information contents. Direct knowledge "from the inside" is never negligible but may be fuzzy when it concerns complex structures of interactions, aggregation across heterogeneous groups, relative importance of conflicting effects or even forms of rationality in individual behaviour. A third source of information then becomes relevant, namely observations collected at the microeconomic level, mainly from individuals, households or firms, either by statistical censuses, or by sample surveys, even in some cases by experiments on groups of students or other subjects.

It is typical for knowledge of a macroeconomic behavioural law to result from many investigations led within an overall framework implying three steps: (i) modellization of what we directly know about the autonomy and objective of the individual agent, (ii) econometric inference, from microeconomic data, about unknown parameters of the model, (iii) aggregate analysis transposing the results of the two first steps, then testing and estimating on macroeconomic data the specification derived from the transposition.

But this convenient scheme well applies only to truly behavioural laws, those concerning clearly identified actions of clearly identified agents, whose autonomy can be unambiguously characterized, whose significant interactions with other agents are well captured through easily observable variables. Macroeconomic theory involves other kinds of relations, which are neither defining identities, nor simple transpositions of budget constraints, nor still market equilibrium equations. Remaining relations may be called "*adjustment laws*". Typical cases are equations applying to changes in the general levels of wage rates or prices. For adjustment laws macroeconomic data provide by far the main source of information, because direct inside knowledge concerning the phenomena involved is particularly fuzzy. Our knowledge for such a case rather is the outcome of a bulk of empirical fits performed on many data sets, for many countries, regions and industries, and concerning different periods. Empirical exploration then is inspired by simple-minded ideas that seem likely to agree with many potential explanations of the changes in question.

4. Why several theories?

Aiming first and foremost at the positive explanation of phenomena, and so forced to accept existing practices, structures, and institutions as they are, macroeconomic theory has a very complex domain to study: a large variety of agents, many operations classified in multiple categories, operations also involving time and uncertainty in an essential way, legal and regulatory systems that vary and evolve…

Under such conditions we should not be surprised to hear that theory has difficulty in getting a tight grasp of phenomena; theory is indeed on the whole hesitant and heterogeneous. It offers incomplete models, which may supplement one another but are not fully incorporated into a consistent grand system. The accuracy of the results leaves much to be desired, and is almost always poorer than would be required for a persuasive explanation and secure applications.

Moreover, on some important issues, conflicting theses exist and are likely to simultaneously long persist before one of them emerges as commonly accepted. From this point of view a deterioration, concerning the core of a large part of macroeconomic theory, occurred since the 1960's, when there was a wide consensus to finding in Keynesian theory the framework for the study of short-run macroeconomic phenomena. It is now common to say that there are various conflicting "paradigms" in macroeconomics and to give them labels, with reference to lists of proponents ready to fight for their respective causes.

Although widespread nowadays, the image is a caricature. If it were not for the temper of some economic theorists and for the impact of the star-system into academia and the media, a more balanced judgement would prevail. It would make the point that, confronted with a very complex world, macroeconomists can now use several theoretical approaches, which are more complementing each other than competing with one another. They are complementary in the same way as are views of a complex object seen from different angles: each view offers a different simplification, but all views are simultaneously useful. Students of macroeconomics should be ready for lucid eclecticism.

The reasons for, and the practical consequences of, the break in the earlier consensus should also be properly judged. There was an empirical reason: the consensus of the 1960s could not adequately cope with the rising rates of inflation and unemployment experienced in the 1970s. There was also a theoretical concern, namely to dig deeper into the microeconomic foundations of an analysis that still owed much to intuition. The practical consequences of the reconsiderations that had to occur were much less dramatic than was often said: applied macroeconomists in business and government did not discard their earlier modes of reasoning and analytical techniques; they simply took the results with a less unreserved confidence; they enlarged the scope of their arguments and analyses. Returning to the above image, we may say that applied macroeconomists better looked around the complex object in order to see whether the view taken from the familiar angle was not missing some important feature.

Now in the 1990s, considering the resistance of applied macroeconomists and the return of academic fashions back to some older concerns, we are reaching a more balanced evaluation of the difficulties of the discipline. Whatever conflict of paradigms still remains, it no longer hides the problems coming from the low degree of mutual cohesion between the various pieces of our knowledge.

5. A pedagogical dilemma

Confronted with the present diversity and low degree of cohesion of these pieces, a textbook aiming at presenting the main parts of the whole field must decide on how to best deal with a delicate dilemma. It must not pretend to offer a tightly articulated general theory into which each fact would find its natural place; indeed, economics did not yet reach the stage where such a theory would exist and be proved reliable. But it must also avoid an excessive eclecticism, which would convey the feeling that anything could be said about macroeconomic phenomena and that anyone could find, in the present scientific corpus about them, the thesis fitting his or her personal tastes or interests of the moment.

The dilemma will be solved here in the following way. Unity of macroeconomic theory ought to appear not in the explanation of the various phenomena examined, but rather through a common conceptual framework and common methods for posing problems and enriching knowledge. Concepts and variables will receive very similar, if not identical, definitions throughout the book; the issues raised will be inspired by a coherent vision of the scientific objectives of macroeconomics and of the types of problems it ought to tackle; the same approach and the same way of combining deductions with inductions will underlie the various analyses discussed.

But those analyses will not perfectly fit with each other; a fortiori they will not result in a well defined whole, which would embrace all of them. On some phenomena two or several alternative models will be studied, with the assumption that different aspects of the same reality are so explained. This will be particularly apparent in the study of elementary behaviours: in each one of the various cases discussed, concerning a type of operation by a type of agent, no attempt will be made to present a unique formalization for the corresponding behavioural law. We shall, however, make an attempt at understanding how alternative formalizations complement each other.

Complexity of phenomena will also impose on our inquiry to separately deal with various aspects of behaviours or interdependences, even when separation will not be better justified than by analytical convenience, i.e. when separability could not be proved to hold except under very exacting and unrealistic hypotheses. No theory can provide more than an approximate representation of the real world; among the approximations made by macroeconomic theory the assumed separability will often be particularly problematic, as well as the assumed level of aggregation.

The reader must therefore know that he or she will not find here *the* theory of macroeconomic phenomena, but rather a set of partial and revisable analyses, all proceeding however from the same methodology. Such is indeed the present state of macroeconomic theories, which display disparate elements for a system still to be fully built.

6. Top–down or bottom-up?

Macroeconomic knowledge is both analytical and systemic, or equivalently synthetic: analytical when it concerns separately the determinants of particular aggregates such as business investment or the rate of increase in wages, systemic when it concerns the global phenomena, economic growth, depression or business fluctuations. Formalized knowledge exists in both cases, for the theory of business investment or wage inflation, as well as for growth theory or business cycle theory. In applied macroeconomics both types of knowledge are relevant; depending on the issue, one type or the other will prove more useful: the issue may focus on the reasons why investment is more sluggish than expected, or it may focus on prediction of a business upturn.

Teaching macroeconomics requires that attention be given to both types of knowledge. A textbook devoted only to the systemic part will convey an incomplete and unbalanced account of present knowledge, even if it provides good references to books and articles dealing with specific topics such as business investment. Modern macroeconomic culture assigns as much importance to the analytic as to the synthetic part.

In favour of priority to, if not exclusivity of, the systemic approach we might argue that behaviour of agents depends on the economic system in which they live. The assertion is unquestionable when meant to apply to the real world; it must, of course, be always remembered in applied macroeconomics. But its implications in theory building and teaching are far from direct. A perfect theory of a particular economic system would provide an exact representation of this system and of the behaviour of agents living in it. But we have no perfect theory of any existing economic system. We rather have a large set of theories, each with an imperfect definition of its actual domain of relevance, as well as with other imperfections. A theory of a particular behaviour has to specify which features and properties of the economic system are taken for granted. A theory of a global phenomenon in a particular economic system has similarly to specify which behavioural characteristics of the agents are assumed to hold. For the latter theory to remain manageable, simplicity of the characteristics is required, and this most often forbids close consistency with known facts, which can be better recognized in a less ambitious theory restricting its scope to explaining a behaviour.

In other words, a book covering both the analytical and the synthetic parts of our knowledge has to follow at places the first approach, at other places the second. There are various solutions for the order according to which approaches alternate. For instance a "top–down strategy" might lead us to start with an extensive study of global phenomena thanks to models of the whole economy; a fuller analysis of behavioural and adjustment laws would come later. The order chosen here will, on the contrary, be inspired mainly by the "bottom-up strategy", except that study of the main adjustment laws will come late in the book, after the theory of economic growth.

On reflection the bottom-up strategy should appear best suited to the uncertain state of our macroeconomic knowledge. It requires that, before erecting embracing systems, we consider how their various building blocks resist the test of close scrutiny. It protects us against the risk of too quickly accepting hypotheses which would facilitate construction of models of the whole economy but later undermine the relevance of these models.

7. Limitations

Important developments of macroeconomic analysis extend to public economics, monetary economics and financial economics. They are then often motivated by the needs of operational studies of economic policies, notably of fiscal or monetary policies. Investigations may be sufficiently elaborate to become subjects of specialized books. A general text like this one cannot cover all such developments, motivated by policy applications. But the bounds within which the author decided to remain here may appear to some readers as severe, more severe than for some extensions concerning other macroeconomic topics.

The bounds result from the decision to stick to a very high level of aggregation in this respect in models of the whole economy. The financial sector is not even identified separately: banks, insurance companies are all assumed away, or aggregated with "government". This is quite common in macroeconomics, but worth remembering. Correlatively, financial operations, much aggregated themselves as will be seen, are supposed to involve only households, non-financial firms and government. No systematic attention is given to the fact that monetary policy often acts through the banking sector. Similarly, no attention is paid to the actual structure of the tax systems; often here taxes do not appear at all; when they do, it is mainly with an exogenously given global tax rate.

This limitation means that the reader should not expect to find in this book a complete introduction to fiscal and monetary policies, not even to their macroeconomic effects since those depend on more detailed characterizations of policies than will be identified. What ought rather to be expected is a set of methods for the analysis of these effects and a set of problems that were raised by macroeconomists in their discussions with policy makers.

A still more disturbing limitation is the choice made here to consider just a closed economy. In practice macroeconomists work in open economies, i.e. in countries trading with others; this may make a great difference: some macroeconomic shocks come from abroad; reactions to domestic exogenous shocks may spill over abroad; they may generate feedbacks; so, the macroeconomic behaviour of an open economy may significantly differ from that of the ideal closed economies studied here; it may differ quantitatively, even in some respects qualitatively. Actually, when considering empirical evidence extracted from macroeconomic data or when referring to experience with macroeconometric models, we shall have at places to reckon with existence or effects of foreign economic relations; we shall then use ad hoc means of doing so.

The choice to limit in principle the scope of this book to a closed economy follows from the consideration that macroeconomic analysis proceeds along essentially the same lines whether it is meant to apply to closed or open economies; new dimensions have to be introduced in the second case, so that arguments and models are more intricate; but a good understanding of the first case makes transposition to the second easy. It would, of course, have been possible to add sections in various chapters so as to show the transposition and its results. But the book is already quite long. Again, priority is here given to methods of analysis and to questions to be answered, rather than to complete coverage in the report of results.

8. The programme

Two long chapters will provide most of the building blocks for models of the whole economy. They will study the behaviour of the two main categories of agents, respectively households, who work, consume and invest their savings (Chapter 2), and firms, which build productive capacities and operate so as to produce and sell goods or services (Chapter 4). But before embarking into these two chapters the reader will find introductory material in Chapter 1, devoted to expose in general terms the elements of the conceptual apparatus of macroeconomics, as well as to elucidate some conventions used in the book. Chapter 3 will examine the formal representation of the technology of production; it will so prepare not only the study of the behaviour of firms, but also the introduction of the productive sector in some models of the economy in which firms have not to be identified as decision makers. Aggregation problems will make their first real appearance in Part 2 of Chapter 2, but will also be discussed later, particularly in Chapters 3 and 4.

How economic growth is analyzed in macroeconomic terms will be seen in Chapters 5 and 6. They will cover a fairly large number of questions, approaches and theories, some of them already posed and discussed many decades ago. From finding out why economies grow to wondering about intertemporal “market failures”, from purely empirical investigations to abstract modelling, from neo-classical theories of “optimal growth” to theories about “the instability of capitalist development”, many themes have to be discussed. Treatment of the broad subject will be organized in two parts focusing on two simple questions. Chapter 5 will ask: how should economies grow? The macroeconomic side of this normative question is essentially simpler than that of the positive question considered in chapter 6: how do economies grow?

The same distinction will not be suitable for structuring the discussion of short-run phenomena and business cycles. The normative aspect does not fully recede; indeed, the search for appropriate economic policies was and remains a dominant motivation for many macroeconomists working on business fluctuations. But this search is intimately linked with progress in understanding the phenomena. Thus it would make little sense to have two separate chapters dealing, respectively, with the positive and normative questions. We shall rather distinguish between static and dynamic approaches. Chapter 7 will be devoted to theories and models arguing in terms of a "short-term equilibrium", supposed to represent the determination of macroeconomic variables in the proximate future. Discussion of the Keynesian theory will be the main subject of the chapter. Giving a large part to the results of econometric investigations, the two following chapters will present what is known about the dynamics of the phenomena and about the potentialities of macroeconomic policies for stabilization. The progression will mainly be from the main short-run trends, in Chapter 8, to larger sets of variables, simultaneously evolving according to more or less periodic fluctuations up to a longer horizon, in Chapter 9.

Chapter 8

The Dynamics of Inflation and Employment

Volume B of this book was devoted to the theory of economic growth (Chapters 5 and 6) and to the short-term macroeconomic equilibrium (Chapter 7). Volume C will now concern all the intermediate range of analysis in which the time dimension will be explicitly present and with a different purpose than just understanding or controlling long-run trends. It may be in order to follow the impact of policies or other exogenous changes during a few subsequent quarters after their occurrence; it may be in order to analyse the longer profiles of business fluctuations.

In any case, the dynamic analyses to be discussed involve the simultaneous evolutions of a more or less large number of variables. They often focus on the evolution of quantities on the one hand, of prices and other nominal variables on the other. For presenting this important material, we shall find it convenient to first limit attention to bare essentials of the simultaneous evolutions. At the expense of neglecting many aspects of business fluctuations, the simplification often is an inescapable pedagogical device, and is indeed made in a very sizable part of the macroeconomic literature. It will be accepted all through this chapter. The next Chapter 9 will have the function of putting the same elements in place into broader and less sketchy modellizations. We shall also defer to the next chapter the study of the cyclical features of business fluctuations. In this one we shall hardly go beyond the consideration of lags in economic relations.

Study of the short-term equilibrium, the central purpose of Chapter 7, mainly aims at finding the level of current activity in the near future. Most often it takes as given the whole past evolution, as well as agents expectations. It considers what impact changes brought to instruments of economic policy will have in an isolated well defined period. Neglecting the temporal dimension of economic phenomena, it ignores important aspects of reality: past evolution matters, not only because of assets and liabilities accumulated by households and firms, but also because of the trends in changes experienced by various magnitudes, trends to which behaviour have adapted; expectations may depend on how current equalization between supplies and demands is being brought about; short-term macroeconomic policy will have effects spreading over several quarters of year and even years, beyond the intended horizon. We recognized the importance of this time dimension in Part 5 of the preceding chapter, but did not then pretend to go into a full dynamic study of its consequences.

Static analysis is particularly inadequate for dealing with inflation, which was twice introduced in Chapter 7. In Section 2.5 we saw how an *ex ante* excess in the demand for goods over supply could be eliminated by a sufficient increase in the price level. But the model we used did not provide a genuine theory of inflation since it did not take into account the fact that any price increase may lead to expectation of further increases, and that this is likely to have an immediate effect on individuals' behaviour. Section 4.4 defined a temporary equilibrium which, in the case of excess demand for goods and for labour could, with sufficient rationing, be achieved with no change whatsoever in prices and wages. Such an equilibrium can provide, at best, one link in a complete theory of inflation. To study inflation more closely we obviously have to resort to models which are, right from the outset, intertemporal and dynamic. This is why we are going to start in this chapter from basic premises, before borrowing from the previous chapter about the way in which the aggregate supply of, and demand for, goods and labour are formed.

But this chapter has a wider spectrum of purposes, which may be now introduced by reference to the history of macroeconomic ideas and research during the second half of the twentieth century. After demand management policies had been defined and studied in a Keynesian spirit during the 1950s, it soon became clear that short-term macroeconomic policy was most often faced with a main dilemma: either to stimulate economic activity and employment, but then to fuel inflation, or to ease pressure of demand on productive capacities, but then to run the risk of increasing unemployment. Characterizing the dilemma became, and since then remained, a major theme for empirical macroeconomics. The "Phillips curve", showing how wage inflation depended on the level of unemployment, was an important tool, which was made more relevant in the early 1960s when adaptive price expectations were also systematically recognized as being determinant for wage inflation.

Already at that time monetarism was offering an alternative framework to Keynesianism for the study of macroeconomic policies. More inclined to stress long-run effects, inspired by the idea that a simple fixed monetary rule would perform better than budgetary activism, it raised a number of valid questions and introduced useful concepts, such as that of rational expectations. For instance it asked: under which conditions will an announced policy appear credible to private agents? Under which conditions will an anticipated policy be effective? Such questions opened the way in the 1970s and later to new theoretical investigations, which were led thanks to more or less persuasive simple models.

By that time drawing conclusions about the effects of policies had become so challenging that serious macroeconomists had to turn their attention to empirical evidence. Part of it came from old monetarist concerns, part from Keynesian econometric fits; but a new econometric approach, the VAR approach, took as its favorite field of experimentation to characterize the effects of monetary policy.

Throughout those decades it had became more and more common to view the macroeconomy as being subject over and over again to random shocks, some of which even possibly arising from policy decisions. The strategic aspects of the behaviours of private agents and public authorities was then stressed.

In order to cover such a wide spectrum the chapter will move to and for, often returning to issues earlier introduced. It will hopefully be organized from the simple to the more complex.

Part 1 will provide a general description and a first analysis of the inflationary process. Without claiming to build a complete model of this process, it will concentrate on various aspects of it, especially on new ones in comparison with those considered in the previous chapters. Hypotheses about price expectations will be discussed. In order to proceed further in several steps, we shall throughout Parts 2 to 5 avoid explicit introduction of the labour market (remember that in Chapter 7 Parts 1 and 2 also avoided the introduction in question).

Part 2 will deal with the role of the pressure of demand for goods on productive capacities and with a first modellization of the main dilemma faced by stabilization policies. Interesting issues about the latter policies were discussed by some authors under an extreme assumption, namely the fiction that public authorities directly controlled the rate of inflation. These issues and discussions will be briefly surveyed in Part 3, where the reader will meet the idea that credibility of public policies matters. The following Part 4, while still accepting great simplifications, will pose the policy problem in terms closer to those used in Chapter 7, with however endogenous price expectations, either adaptive or rational. The distinction between instrument rules and target rules will be introduced. Part 5 will be devoted to discuss the effectiveness of monetary policies, providing empirical tests of monetarist theses. The VAR techniques will be defined and their application to the issue be surveyed. Reference will also be made to less formalized inquiries into the actual strategies of monetary authorities.

Part 6 will deal with the reciprocal short-run relationships between the goods market and the labour market, particularly with what was learned by econometric research about these relationships and with tools such as the NAIRU derived from the knowledge so acquired for macroeconomic diagnosis and policy analysis. Part 7 will discuss, in this enlarged framework, aggregate demand management policies, and briefly price and income policies. It will show why the relative importances of real and nominal shocks matter for the choice of monetary policy rules.

Models presented, particularly in Part 6, try to be realistic and are indeed closely inspired by results of econometric research. Nevertheless, they might seem to have a rather weak basis, since their microeconomic foundations are not so clear. The difficulty comes from the fact that the links between individual behaviour and aggregate outcomes are particularly complex here: "adjustment laws" included in macroeconomic theories of inflation and of its control represent the effects of numerous microeconomic decisions in which the choice variables are not always those seen at the aggregate level. Thus, it is very difficult to extricate the microeconomic foundations of these laws. Recent research has, however, devoted many efforts to attack the problem from various sides. The aim of Part 8 is to give an account of the results of this research, elements of which were already introduced earlier in this book.

1 Inflation as a Process

To talk about inflation is to speak of a continuing increase in the general level of prices and wages. The sequence of phenomena which result from such an increase, which accompany it and sustain it, are well known. The various prices and wage rates are raised successively, each rise supplying the justification for the following one. As soon as inflation is sizable, it is a process fuelled by its own momentum; it is then sometimes called the "inflationary spiral".

A theory of inflation aims at singling out those factors which can influence the path of this process. In order to construct such a theory we must first of all decide with which main empirical facts we ought to deal (Section 1.1). Then we must sort out those factors which accelerate or slow down the rate of inflation (Section 1.2). The role of increases in the money supply must be considered, even in our initial exploration of the subject (Section 1.3). It is also useful to clarify ideas about the role and the nature of expectations, which strongly influence agents' behaviour (Sections 1.4, 1.5 and 1.6); also important is the clarification of ideas about fundamental trends in the evolution of interest rates and asset prices over time (Sections 1.7 and 1.8). Finally, the chapter would be incomplete if we did not elucidate, from the outset, the reasons why governments should fight against inflation (Section 1.9).

1.1. Costs, prices and wages: the spiral

The main aspects of the way in which prices are actually determined are well known. We talked about them in the previous chapter when we had to examine the reasons why price flexibility, which assumes that the law of supply and demand functions fast and fully, does not hold in the short run. We must take these aspects up again whenever we are describing the inflationary process.

1. Fresh food products are almost the only case where prices are fixed from day to day, in order for current demand to absorb the amounts supplied. For some products or specific deliveries, as well as for the provision of specialized services, the price is negotiated between seller and buyer; but this does not happen very frequently. Usually, the price of the good or service is fixed in practice for a certain length of time. The decision to change it is normally made by the producer or the seller, sometimes by public authorities. It is therefore useful to understand the motives for changes in prices, and the factors which determine amounts of those changes.

The usual reason put forward to justify a rise in prices and to explain the exact amount stems from an increase in costs: the manufacturing and transportation costs for a producer, the purchase price and the cost of distribution for the seller. It is said that because costs have risen by 5% then the price has to rise by 5%. Successive price increases would reflect quite simply the necessary transmission of an inflation whose origin should be looked for in the past.

However, we do have to look at things more closely, for the transmission process is not absolutely rigid, even in case of rapid inflation. Each increase can be made earlier or later. It will be made without hesitation if demand is buoyant, whereas the seller would certainly think twice about it if clients were already starting to reduce their purchases, or to look for other suppliers. It would be made quickly if the increase in costs were expected to continue; on the other hand, it would often be put off if there were a possibility that it was exceptional and not likely to last long.

An increase in prices can be the occasion for adding a little extra margin or, on the contrary, for trying to "squeeze prices" as much as possible. The decision will be oriented in one of these directions depending upon the state of the market and the state of expectations about costs and pricing policies of competitors. Observing a high pressure of demand would lead to a generous calculation of new prices, as would the expectation of substantial pressure in the future or of large increases in costs. Similar considerations apply when the general price level hardly changes and people do not have to take an overall rate of inflation into account.

The case of public tariffs is hardly any different. It is similar in that cost increases must sooner or later be passed on; dates and amounts of the increases also depend upon expectations about the future development of costs over time. However, decisions are less sensitive to the state of the market and are more sensitive to the aims of economic policy. Sometimes concern for balancing public budgets will lead to a search for “true prices”; at other times fight against inflation will be used as a justification for keeping tariffs stable in spite of an increase in costs.

2. The main reason given for a wage increase, and for the rate of that increase, is a rise in the “cost of living”. This reason is paralled to the one invoked for price increases and completes the cycle, typical of the spiral from costs to prices, from prices to wages and from wages to costs. Nevertheless, repercussion of price increases is not completely rigid. It would be so if statutory indexation existed and were strictly applied. Adopting an explicit and general indexation of wages requires a political decision, which may seem natural in countries where inflation is rapid and long lasting but also takes many other considerations into account. Without legal indexation, collective agreements and practices also tend to establish a direct link between consumer prices and wages; the more sensitive the country is to inflation, the closer the link, adjustments being more frequent and more automatic. In Part 6 we shall come back to the econometric measure of the repercussion of prices on wages and we shall see that it is not usually immediate, and is often incomplete.

The change in the cost of living is not the only thing that matters. In the long run we observe that the rise in real wages does not much differ from that in the productivity of labour. But in the short run the link is weak. Not only is the development over time of real wages more regular than that of labour productivity, but wage increases also depend on changes in other costs, on the state of the market, on the state of expectations and on legal provisions and rules. It is not surprising to find that, at a given rate of consumer price increase, wages progress particularly fast when firms experience difficulty in finding the manpower they are looking for, when everyone expects rapid increases in prices and wages, when raw materials are cheap and financial charges light, or when there is a sharp rise in the minimum legal wage; the contrary is obviously observed in the opposite circumstances.

3. The inflationary spiral thus appears to be a process by which nominal increases are transmitted, but a process in which there is a certain amount of play. It is precisely on this play that we should concentrate either in a precise theory of inflation or in a theory of the short-run dynamics of a non-inflationary economy. But, of course, any model of inflation must first of all show the general process of transmission.

Before going any further we should also be aware of two special features of the phenomena we are going to examine. In the first place, the path of inflation over time has the peculiarity of being explained in part by its own past, and in part by what it is expected to be in the future. Previous increases in costs and prices are passed on, but with lags which give some inertia to the process. Each time an increase is decided upon, new prices or new wage rates are adopted for a future period, as a function of what costs and prices are expected to be during the period. But once the process is started, it quickly becomes impossible to make the distinction, for each increase, between what is the consequence of former increases and what is due to expected future increases. We shall indeed see that trying to make such a distinction can lead to insuperable difficulties. In particular specific conventions will be required for defining "autonomous" factors affecting expectations.

In the second place, the process is asymmetric. We saw this in the previous chapter when we looked at the reasons why price rigidity was particularly noticeable in the case of excess supply, when competition alone should lead to a decrease. We there ignored the possible existence of a general inflationary trend. If we now take this into account, we have to conclude, on the basis of arguments then given, that excess demand accelerates inflation much more than the same amount of excess supply would slow inflation down. Without trying to be precise, we can say that inflation would have a spontaneous tendency to accelerate if the motion were not otherwise barred; accidental shocks happen which sometimes lead to an increase, sometimes to a decrease in prices; but because the results of these are asymmetric, accelerations tend to be more noticeable than decelerations. If the fight against inflation so often belongs to the objectives of economic policy it is precisely because it seems that deliberate action is required to compensate for the spontaneous tendency of inflation to accelerate. The force of the tendency however seems to vary, being at some times stronger than at others for reasons which are not easily identified.

1.2. Cost-push inflation and demand-pull inflation

1. Since the inflationary process is not totally rigid and since its development over time is influenced by various factors, it is useful to classify these factors according to their characteristics. With this in mind, we sometimes talk about "cost-push inflation" and sometimes about "demand-pull inflation". We mean by this that the origin of inflation is found in costs in the first case, and in the pressure of demand in the second case.

These terms suggest first of all a transition from a situation of price stability to a situation of inflation. They can obviously be applied more generally, as soon as we recognize that we are not referring to the very beginning of a process, which has often continued over a long period of time, but rather characterizing what was responsible for its recent acceleration. Thus an economist, when analysing business trends, will talk about cost-push inflation or demand-pull inflation according to which he or she considers to have been the most important factor in changing the rate of price increase during the period considered. Notice also that the reverse terms to cost-push inflation and demand-pull inflation are cost deflation and demand deflation.

The purest case of cost-push inflation is when the "cost push" comes directly from abroad: an increase in the price of oil or imported raw materials, and which is higher than the increase in the general price level. The same term has often been used for periods in which increases in wages have been exceptionally fast, or even, during periods when other charges on firms have risen sharply (taxes, financial costs, etc.). Pure demand-pull inflation, on the contrary, characterizes a situation where cost conditions do not change abnormally, but where there is substantial excess demand on the goods market and on the labour market, so that the price and wage increases decided by economic agents are calculated with a generous margin.

In reality things are often less clear-cut. Many factors may change earlier trends, some acting on costs, others on the supply or demand in various markets. But, building models, we can keep the idea of separating out on the one hand, autonomous influences on costs, or even prices, and on the other hand, influences which stem from the pressure of demand on supply.

2. In this chapter it will be convenient to use the notation g_{pt}, or even simply g_t, to denote the rate of increase in the price level between period t and period $t+1$:

$$g_t = g_{pt} = \frac{p_{t+1} - p_t}{p_t}. \tag{1}$$

In the study of this rate, various closely similar formalizations will be used, the following one, for example:

$$g_t = g_t^e + z_t + \pi(d_t), \tag{2}$$

where g_t^e is meant to characterize the "expected" rate of price increase, or more generally, the basic instantaneous trend of this increase, z_t the impact of autonomous influences on prices or costs, π an increasing function and d_t a measure of the pressure of demand.

An equation such as (2) thus distinguishes three terms which can be analysed separately. It may be used when we are diagnosing inflation trends; we may then speak of cost-push inflation if the second term z_t is positive, and of demand-pull inflation if the third one $\pi(d_t)$ is positive. The equation may also enter into a theoretical model. It then formalizes an "adjustment law" (at this stage its microeconomic and econometric foundations may seem rather obscure; we shall come back to them later).

Actually, the term "adjustment" is used here in a slightly different sense from that used in the previous chapter. There the purpose was to characterize the adjustments leading from *ex ante* magnitudes to the short-term equilibrium; we assumed those adjustments to be instantaneous. Now we are concerned with changes from one period to the next, that is, from one temporary equilibrium to the next. The distinction is particularly clear if we assume that the price level p_t of period t is not dependent on the economic conditions in period t, but only on earlier conditions. So, an expression like Equation (2) characterizes how conditions in period t determine the change in the price level between period t and the next one.

Although the two types of adjustments are formally different, the economic phenomena involved are the same; justifications given to different formalizations are then often similar. In particular, some models assume that production adjusts instantaneously to demand, as was the case in the previous chapter whenever we discussed the Keynesian equilibrium. Other models may admit that production is decided one period in advance and that a gap remains in period t between intended and actual sales, because of involuntary accumulation or decumulation of inventories; from period t to the following one, production may then be assumed to change, so that the involuntary part of the change in inventories in period t would be exactly eliminated in period $t + 1$ if actual sales in that period were exactly equal to intended sales (we shall examine such models later). The adjustment of production to demand is so assumed to act instantaneously in models of the first type, and to act with a lag in models of the second type; but except for that, we may have two similar models of the same actual process.

1.3. Money and inflation

1. A theoretical treatment of inflation that ignored monetary phenomena would be open to criticism, partly because it is customary to associate the increase in the price level (the definition chosen here of the term "inflation") with too large a supply of money (the initial definition of the term), partly because there remains in the collective awareness of economists the memory of inflationary experiments with exaggerated monetization (especially that of Law at the beginning of the eighteenth century). Now, this chapter seldom assumes a direct link between increases in the supply of money and changes in the price level.

It is essential to understand right from the beginning why the assumption is not imperative. In order to show this we shall start from the thesis of a direct link, recognize its partial truth, but also show why it cannot strictly hold when applied to the short run.

We know that there are various simple expressions for the quantity theory of money, for example, the "Cambridge equation";

$$M = kpy, \tag{3}$$

where k is a fixed coefficient linking the demand for money to the value of aggregate output. The argument then runs about as follows: if coefficient k and the volume y of production can be taken as exogenous and if the demand for money has to adjust to the supply of it, then any assumed change in the money supply M must result in a proportional change in the price level.

This argument is fairly persuasive when it concerns the long run. We used a similar reasoning in Chapter 5 when we were considering the choice of financial strategies for growth. Actually we then recognized that, along a proportional growth path, output may depend on the real interest rate, itself linked to capital intensity and to the productivity of labour; we recognized that the coefficient k may depend just as well on the nominal interest rate, as on the ratio between the value of private wealth and the value of current income. But we had to take into account that, in proportional growth, the rate of inflation had to be equal to the difference between the growth rates of the money supply and of the volume of production respectively.

Even for long-term reactions we must nevertheless understand that the link is less simplistic than some applications of the quantity theory of money suppose. It is true that if, from some time on, the growth rate of the money supply is sharply lowered and then maintained constant for a long period, the growth rate of the value of production must in the long run fall and approach a value close to that of the money supply. But this does not imply that the initial proportion between the money supply and the price level is then re-established. In the notation of Chapter 5, Part 6, this proportion was denoted by μ. Now, not only does the liquidity ratio μ change with the rate of inflation, but output can be significantly affected by economic developments in the intermediate period and by the final changes in variables such as the real interest rate and the rate of taxation[1].

[1] We must also take into account that the money supply is not directly controlled by the monetary authorities and that even the conditions which govern this control can be affected in the long run by economic conditions. In a modern state, financial structures were created and evolve so as facilitate complex economic operations. A long period of restriction of liquidity can stimulate the

2. However, the essential thing here is to observe that the intermediate adjustment period cannot be short and should not be ignored in any theory of economic evolution that is not limited to studying long-run growth. It is common place, but nevertheless useful, to remind ourselves that no economic agent directly takes into account changes in the aggregate money supply when he or she fixes the prices or wage rates that are under his or her control. The relationship between the money supply and the price level is therefore an indirect one; it works through a set of progressive adaptations, which precisely we must try to understand.

Now, the sort of decomposition made in Equation (2) is helpful for tracing impacts of the monetary factors of inflation. The most directly perceptible response of changes in prices to changes in the money supply happens through the pressure of demand, represented by d_t in Equation (2).

In the previous chapter we saw how a restrictive monetary policy reduces the demand for goods and services. An increase in the interest rate, which is necessary in order to allow the redistribution of assets towards non-monetary forms, depresses investment, inventories accumulation and purchase of consumer durables. This effect is reinforced by any credit rationing that banks, or even monetary authorities, may impose to accompany the reduced money supply. In all sectors where production is limited by a lack of demand, this fall in demand for goods and services produces, in particular, a drop in the demand for labour and the multiplier effect comes into play.

We can well understand the difficulty of disinflation if we note that the function $\pi(d_t)$ has most often a low elasticity: the pressure of demand must be substantially reduced for a fairly long period if one wants to achieve a substantial fall in the rate of inflation. We shall deal more specifically with this in Parts 4 and 5 of this chapter. At this stage let us note that the low elasticity of $\pi(d_t)$ is precisely the reason why the quantity theory of money turns out to be inadequate for a proper discussion of all problems raised by the fight against inflation.

Since the term z_t of Equation (2) concerns, in particular, autonomous increases in costs, introducing a restrictive monetary policy may even have a small inflationary effect which, to start with, reduces the impact of the change in $\pi(d_t)$: a rise in the interest rate immediately increases the financial costs of firms, which are on average net debtors.

We shall see later on that there is a great deal of inertia about the "basic instantaneous trend" g_t^e, and that it most often tracks with a lag a moving average of the observed rate of price increase. It does not normally play an active

appearance of substitutes for money and for short-term credit, or the development of practices which in fact make some assets liquid which were not so previously.

role in any favourable or unfavourable change in the rate of inflation. However, since this "trend" is affected in particular by agents' expectations, one can well imagine circumstances in which government policy would succeed in engineering a real break with the past, and thus provoke an immediate fall in g_t^e. History shows that "monetary reforms" have at times thus attained their objective, especially after periods of hyperinflation. But history also shows that most government efforts to alter expectations, in any lasting way, have failed and that, where they have succeeded, the reform was much more than a simple technical monetary operation.

So, we may conclude that the approach suggested in the previous section should allow a good analysis of the effects of monetary policy. We should certainly check that the model eventually chosen provides the required consistency, in the long run, between the growth of the money supply and the rate of inflation. But this is only one requirement amongst many others.

1.4. Simple formulas for price expectations

The most natural way to exhibit the inertia in the basic instantaneous trend g_t^e is to assume that its rate is determined from the past history of prices by some stable formula such as:

$$g_t^e = \sum_{\tau=1}^{\infty} b_\tau g_{t-\tau}, \tag{4}$$

where the coefficients b_τ are fixed numbers[2]. It is also natural to assume that, if the rate of inflation has been constant in the past, then the trend should increase exactly at this rate. Thus the sum of the b_τ ought to be equal to 1:

$$\sum_{\tau=1}^{\infty} b_\tau = 1. \tag{5}$$

Later on we shall frequently use an expression like (4), and even more specifically, a special case of this expression:

$$g_t^e = (1-b) \sum_{\tau=1}^{\infty} b^{\tau-1} g_{t-\tau}. \tag{6}$$

[2] Here and all along this chapter, we shall often consider infinite sums such as this one. Their convergence will be assumed, which will never be a serious restriction in the context in which these sums will be used.

The number b, which is positive and smaller than 1, thus characterizes the inertia in the trend: the higher this number, the more important are the early values of the rate of inflation in determining the present trend.

Accepting an equation like (4) or (6) does not necessarily mean that we take a position as to exactly which reasons explain the existence of a basic trend in inflation. The equation establishes a simple "mechanical" relation between the past and the present. Many theories might well lead to the conclusion that such a relation will, most often, provide a good first approximation of the basic trend in inflation. For example, this would be true for a theory in which cost increases would be passed on with a lag at each stage in the chain of industrial transformations and commercial distributions through which products have to pass, whether made for consumption or investment.

But we must concentrate particularly on the role that the formation of price expectations can play in determining the trend. This role is bound to be important, as we saw in Section 1.1. So, we are going to argue, most often in what follows, as if the term for the trend g_t^e would correspond exactly to the expected rate of price increase. The main question will then be to see whether expressions such as (4) or (6) correctly represent price expectations.

Simple formulas can follow from ideas about the formation of expectations. We may in particular think that, in normal times, a fairly natural and widespread behaviour leads to "adaptive expectations": agents revise their expectations as a function of the errors that they observe in their past forecasts. Revision would be immediate, for example, with the following expression, which is often suggested:

$$g_t^e = g_{t-1}^e + (1-b)\big(g_{t-1} - g_{t-1}^e\big). \tag{7}$$

Here the number b is smaller than 1 and positive. The closer b is to 1, the less agents allow themselves to be influenced by the last error observed. There is a close link between the "extrapolative expectations" obeying formulas such as (4) and adaptive expectations. Indeed, it follows from (7), for instance, that expectations obey also Equation (6). More generally, adaptive expectations are not distinguished from extrapolative expectations, except when their justifications are formulated.

There are reasons to think that a formula such as (7), or equivalently (6), is too simple for correctly representing the formation of expectations. First of all, the expected rate of inflation for the near future (the next three months, for instance) does not need to coincide with that expected for a more distant future (the next ten years). In some cases there are reasons to believe in a substantial difference between the two. All depends on the views of agents about the process actually governing the rate of inflation.

For instance, we shall see in a moment that applying Equation (6) is rational for agents believing that this process is well represented by:

$$g_t = (1-b)\sum_{\tau=1}^{\infty} b^{\tau-1} g_{t-\tau} + \varepsilon_t, \tag{8}$$

where the unobserved shocks ε_t follow a pure white noise process with a zero mean. But econometric studies suggest that it would be simplifying things a great deal to assume that such an autoregressive expression perfectly represents the views of agents on how prices will evolve[3].

A more elaborate intellectual model would apply to individuals holding fairly precise ideas about the rate of inflation that will prevail in the long run. Their short-term forecasts would then be influenced by recent developments, but also by the idea of a tendency towards the "normal long-run" rate $\hat{g}$. In order to formalize such a case we might replace (8) by:

$$g_t = a(1-b)\sum_{\pi=1}^{\infty} b^{\tau-1} g_{t-\tau} + (1-a)\hat{g} + \varepsilon_t, \tag{9}$$

where a would be a positive number less than 1. According to this last model, the rate g_t follows a stationary autoregressive process which has exactly $\hat{g}$ as its mean (process (8) is not stationary). In this case, the rationally expected rate g_t^e is given not by (6) but by the expression found for $g_t - \varepsilon_t$ in Equation (9). Expectations conforming to such an expression are said to be "regressive". In this connection it is interesting to note that econometric analyses of short-term price expectations fitted to an equation such (4) usually require addition of a constant term and lead to estimates of coefficients b_τ whose sum is smaller than 1, as would be the case if (9) applied.

However, we should not be led too hastily to the conclusion that a theory starting from Equations (2) and (4) should reject Equation (5) and assume that the sum of the b_τ is less than 1. In fact, the long-term trend may not itself be constant, nor independent of past history. A more realistic expression than (9) would replace $\hat{g}$ by g_t^1. If this last variable itself followed an autoregressive

[3] See on this subject the analysis of forecasts given by about a hundred experts, E.J. Kane and B.G. Malkiel, Autoregressive and non-autoregressive elements in cross-section forecasts of inflation, *Econometrica* (January 1976).

process like (8), although with a coefficient b_1 much closer to 1 than b is, we would come back to expression (4) with:

$$b_\tau = a(1-b)b^{\tau-1} + (1-a)(1-b_1)b_1^{\tau-1}, \tag{10}$$

whose sum is certainly equal to 1. To sum up: to see whether we should keep constraint (5) we have to consider what period the theory is supposed to explain. If the period is short, we could include in the model a normal long-run rate $\hat{g}$ and take it as being exogenous. This would be much more dangerous in the case of models aiming to deal with medium- or long-run problems.

But the linearity of expression (4) can also be questioned. It does not allow us to take into account that the sensitivity of expectations to recent changes in prices might increase when the rate of inflation itself increases. Yet, it seems to be precisely the case that the higher the rate of inflation the quicker the expected price increase adapts to the actual increase. In order to take this fact into account, we ought to substitute for Equation (7), or equivalently for (6), a similar equation, but in which b, instead of being constant, would be a decreasing function of past values $g_{t-\tau}$ of the rate of inflation. The study of hyperinflations has indeed led to this recommendation[4].

For simplicity we shall consider only linear equations here when we shall incorporate price expectations formation within more complete models. But, we must remember that the values to be chosen for coefficients in these equations must depend on the context studied.

1.5. Rational expectations

1. The expressions we have just been considering can also be accused of giving too much weight to the past in a model of behaviour which directly concerns the future. In order to evaluate the relevance of this criticism we have, first, to well understand the logic of the formation of expectations, second, to study its consequences, third, to turn to econometric results testing various hypotheses about the determination of expectations. Let us then devote this section to a closer look at the hypothesis of rational expectations, whose logical underpinnings are particularly clear.

As we recognized in Chapter 1, expectations are not certain. Except in rare cases, they ought to be represented by random variables; but in order to simplify the analysis, we most often stick to sure variables in our formulations.

[4] See P. Cagan, The monetary dynamics of hyperinflation, in: M. Friedman (ed.), *Studies in the Quantity Theory of Money* (University of Chicago Press, 1956).

These variables can be considered as central characteristics of the probability distributions of the corresponding random variables. We shall assume that the most relevant central characteristic is the mathematical expectation. This would be the case if models were linear and if losses resulting from erroneous forecasts were quadratic functions of the errors (the case in which the "certainty equivalence property" holds). But in many cases it will be hardly more than a convenient notational convention.

So, for expectations as to the rate of inflation, we can write:

$$g_t^e = E(g_t) \tag{11}$$

or even:

$$g_t^e = E(g_t/J_t) \tag{12}$$

if we intend to insist in the notation on the fact that the probability distribution which agents attribute at time t to g_t depends on the set J_t of informations they have at the beginning of period t. These informations concern past values of relevant variables, such as $p_{t-\tau}$ (for $\tau \geqslant 0$) or even $z_{t-\tau}$ and $d_{t-\tau}$ (for $\tau > 0$) in our model; they also concern what agents know about the process generating the evolution of the price level, a knowledge that they naturally ought to take into account in forming their expectations. Informations may also concern new events, such as the threat of a war, the announce of a bad harvest or a change of government. The last feature is potentially a definite advantage of specification (12) over the mechanical formulas of the previous section. Moreover, since agents might take all available informations into account when forming their expectations about the probability distribution of the inflation rate, it is natural to speak of "rational expectations" in characterizing the result of (12).

The compact notation (12) however draws our attention to a first difficulty, because common sense suggests that the set J_t and the probability distribution of the expectation derived from it depend on the agents concerned. Diffusion and interpretation of information are not perfect and unevenly reach various people. Aggregation difficulties are therefore waved away in aggregate models where a single variable g_p^e is used for *the* expected rate of inflation. As we realized, this is a feature of all models derived from analysis of the behaviour of a representative agent; but it is worth noting here in passing in a new framework.

In the practice of macroeconomic theory a more disturbing feature occurs. Indeed, the specification usually translates (12) into what the assumed model should imply about it. We shall see in a moment how the practice applies. At this point let us express it in simple words, which will moreover explain why some economists prefer to speak of "*self-fulfilling expectations*", rather than of rational expectations. It is still better to speak of "*model-consistent expectations*", because the phrase better exhibits the abstract nature of the concept. The theorist builds a model for the representation of the process generating changes in the rate of inflation. He or she then assumes that private agents know the model and believe it is right, hence that they form their expectations in accordance with the truth of the model.

In other words, within the hypothesis that the model of the theorist is correct in its representation of the generating process, is incorporated the hypothesis that agents know this model and accept it as providing a reliable representation, to be rationally used in forming their expectations. The hypothesis involves a circularity about which we shall presently comment. But, above all, it places a heavy responsibility on the shoulders of the theorist claiming that his or her model should be accepted as realistic; it cries out for validation by an econometric test.

Since the formation of expectations are meant to depend on the process ruling inflation and since the process involves these expectations, any model of the phenomenon has the built-in circularity just mentioned. This can lead to properties of the process being different from those we might intuitively think of. A serious study of the analytics of the model and of its solution is therefore called for. An example will illustrate how such a study can proceed and to what it may lead.

The hypothesis of perfect rationality, as spelt out here, may appear too extreme to be true. Models based on it are likely to exhibit more circularity than the phenomena they claim to explain. The idea must be kept in mind although its implications are obscure, because some circularity is likely to remain, even with a more realistic imperfect rationality, as soon as all aspects of the forward-looking nature of expectations are taken into account.

2. A first introduction to the analytics of rational expectation models must focus on a simple case in which there is just one endogenous variable, such as the price level. But the discussion would hide an important part of the difficulties if there was not also an exogenous variable, about which prior independent expectations may have to be formed.

We might consider here a model that was used in the literature dealing with the effectiveness of monetary policy[5]. But we shall not do so because the model

[5] This model seems to have been introduced by T. Sargent and N. Wallace, The stability of models of money and growth with perfect foresight, *Econometrica* (November 1973). Note that the title refers to growth models. But the article is ambiguous on this subject; it has often been quoted in studies on short-term monetary policy. This is when model (13) of the inflationary process becomes particularly inexact.

is obviously inexact. Indeed, it takes, as ruling the process determining the price level p_t, the following equation:

$$\log \frac{M_t}{p_t} = \alpha g_t^e + m_0, \tag{13}$$

where the money supply M_t is exogenous and the coefficient α is negative (the real value of desired cash balances is decreasing function of the expected rate of inflation). We saw in Section 1.3 why such an equation is not appropriate when we want to model how the price level is determined in the short run.

So, we shall consider instead an example which could be applied to pure cost-push inflation (an extreme case and, because of this, not very realistic). We shall assume that some of the costs taken into account in period t are exogenous and represented by the variable w_t, whilst others are a direct function of the expected price level for period $t+1$, that is p_{t+1}^e (this expected level being assumed to play a direct part in the way in which some prices and wages are determined). So there will be, *a priori*, no inertia in the inflationary process; this is certainly inexact but convenient (we shall later relax this assumption). It will thus be possible to use, for the determination of the price level, the weighted average[6]:

$$p_t = \beta p_{t+1}^e + (1-\beta) w_t, \tag{14}$$

where β is a positive number less than 1.

The hypothesis of rational expectations consists, on the one hand, in setting:

$$p_{t+1}^e = E(p_{t+1}/J_t) \tag{15}$$

and on the other, in assuming that agents' information J_t includes the knowledge of: Equation (14), the value of β, all the past and present values $p_{t-\tau}$ and $w_{t-\tau}$ ($\tau = 0, 1, \ldots$). The random component comes from the fact that agents do not know perfectly the future values $p_{t+\tau}$ and $w_{t+\tau}$ ($\tau = 1, 2, \ldots$); they treat them therefore as random variables.

Before going any further, note that the system defined by Equations (14) and (15), where the variables p_t and w_t are random, could be used in a wider context than just pure cost-push inflation. The crucial point is that it assumes w_t to be exogenous. This variable could, in particular, reflect the pressure of demand

[6] Some readers may be bothered in what follows by the natural non-negativity of prices and costs. They may then prefer to think of p_t and w_t as being the logarithms of respectively the price and the exogenous cost.

d_t if this could be assumed to be exogenous. This would often be an acceptable assumption to express the actual state of agents' information; it would be less acceptable to represent the process of price formation (in Section 2.4 we shall discuss how d_t depends on the price level p_t).

The circularity comes here from the fact that p^e_{t+1} features in Equation (14), which represents the stochastic process ruling p_t, whereas this process determines, according to Equation (15), the value of p^e_{t+1}. Note that these two equations imply the following recursive relation:

$$E_t p_{t+\theta} = \beta E_t p_{t+\theta+1} + (1-\beta) E_t w_{t+\theta}, \tag{16}$$

where the notation E_t replace $E(./J_t)$, for convenience in the equations. This recursive relation holds for all non-negative values of θ. Conversely, when the definition (15) is taken into account any process $\{p_t\}$ fulfilling this recursive equation also fulfils (14).

But the system of Equations (16) is not sufficient to determine, uniquely[7], the process followed by p_t from that followed by w_t. We immediately see, that if p_t is a solution then $p_t + \varepsilon_t$ is also a solution, as soon as $E_t \varepsilon_{t+\theta} = \beta^{-\theta} \varepsilon_t$ for $\theta \geqslant 0$ (the random process $\beta^t \varepsilon_t$ is thus said to be a "martingale", i.e. a process $\{\eta_t\}$ in which the mathematical expectation at time t of future values is equal to η_t); as far as expectations are concerned, the solution p^e_{t+1} is replaced by $p^e_{t+1} + \beta^{-1}\varepsilon_t$. The model ought to therefore specify which solution is meant to hold.

The most natural way to do this is to give preference to the special solution called the "forward solution":

$$p_t = (1-\beta) \sum_{\tau=0}^{\infty} \beta^\tau E_t w_{t+\tau} \tag{17}$$

which leads to:

$$p^e_{t+1} = (1-\beta) \sum_{\tau=0}^{\infty} \beta^\tau E_t w_{t+\tau+1}. \tag{18}$$

(Here we use the fact that $E_t E_{t+1} = E_t$ because J_{t+1} contains J_t: it also contains p_{t+1} and w_{t+1}. Moreover, we are assuming that the exogenous infinite sum in the right-hand member of (17) converges.) In this forward solution the

[7] See C. Gouriéroux, J.-J. Laffont and A. Monfort, Rational expectations in dynamic linear models, *Econometrica* **50** (1982) 409–425.

expected value of the price level is a weighted average of the expected future values of exogenous costs; the weights follow a geometrical progression, decreasing to zero the further into the future one goes.

From a purely formal viewpoint we note that the system of Equations (16) would also admit the following "backward solution" if it were meaningful:

$$p_t = -(1-\beta)\sum_{\tau=1}^{\infty} \beta^{-\tau} w_{t-\tau}. \tag{19}$$

Since we have assumed that β is positive and less than 1, this solution is not acceptable: not only is convergence of the infinite sum doubtful, but also p_t and the w_t cannot all be positive. We can easily realize, however, that in other models a solution of this type, where the endogenous variable depends only on past values of the exogenous variables, could have a real significance, no matter how strange this may seem for a theory insisting on the role of expectations (an exact forecast would thus be found, in which agents would stick to a "mechanical" formula, like those considered in the previous section).

3. Let us complete the discussion by considering two reasons for giving weight to past values of the price level in the formation of expectations about it, contrary to what is implied by Equation (18). First it can be considered that expression (14) takes too little account of the fact that, in reality, the price level of each period is affected by adjustment lags in some prices or wage rates. But nothing prevents us from substituting for (14) a slightly more complex equation, for example:

$$p_t = a_1 p_{t+1}^e + a_2 w_t + a_3 p_{t-1} \tag{20}$$

in which the last term reflects adjustment lags for prices (a_1, a_2, a_3 are positive coefficients, whose sum can be taken equal to 1).

To deal with the system made of (15) and (20), we just have to consider:

$$x_t = p_t - b_3 p_{t-1} \tag{21}$$

and to check that:

$$x_t = b_1 x_{t+1}^e + b_2 w_t \tag{22}$$

if $b_1 = ka_1,\ b_2 = ka_2,\ b_3 = ka_3$ and if the number k satisfies:

$$a_1 a_3 k^2 - k + 1 = 0. \tag{23}$$

We may also check that, since a_1 and a_3 are both positive and have a sum less than 1, the two following possibilities for the formation of the price level should be considered.

The smaller solution of (23) leads to $b_1 < 1$ and $b_3 < 1$. Noting the similarity between (14) and (22) we can thus transpose (17) and obtain:

$$p_t = b_3 p_{t-1} + b_2 \sum_{\tau=0}^{\infty} b_1^{\tau} E_t w_{t+\tau}. \tag{24}$$

Note that, since $a_1 + a_2 + a_3 = 1$, $1 - b_3$ equals $b_2/(1 - b_1)$, as it should.

The larger solution of (23) leads to $b_3 > 1$ and $b_1 > 1$, which may be taken as incompatible with the convergence of the infinite sum in (24). Transposing (19) leads to a result that seems mathematically acceptable:

$$p_t = b_3 p_{t-1} - b_2 \sum_{\tau=1}^{\infty} b_1^{-\tau} w_{t-\tau}. \tag{25}$$

But economically, the nature of the dependence of prices on costs, implied by such an equation, seems very unlikely. This solution should be ruled out.

It should also be noted that equations similar to (20) but with more lags, for the determination of p_t, would lead to solutions of the same type as (24), but that more lagged values of the price level would appear. So, we could progressively come back to expressions close to those given by Equations (2) and (4), where it would now be understood that component z_t ought to include expectations as to the future path of exogenous costs.

Another reason for giving weight to past values in the formation of expectations is to recognize that some agents have less information than is assumed in strict applications of the rational expectation methodology. Some agents may not know the "true" model, (14) or (20), of determination of the price level. They may have more naive models naturally leading to extrapolative expectations[8].

We briefly mentioned and used the idea at the end of the previous section when introducing models (8) and (9). Let us elaborate it just a bit more now, considering the case of an agent who, for lack of better knowledge, would

[8] Indeed, any in a large set of formulas for series of expectations may be consistent with optimal forecasts of the future variables under certain special circumstances. See R. Shiller, Rational expectations and the dynamic structure of rational expectation models: a critical survey, *Journal of Monetary Economics* **4** (1978) 1–44.

consider that the rate of inflation will differ in period t from what it was in period $t-1$ by a shock γ_t:

$$g_t = g_{t-1} + \gamma_t. \tag{26}$$

But the shock would not be purely random; it would be more likely positive than negative in period t if it was positive in period $t-1$: observation of acceleration in period $t-1$ is most often followed by observation of acceleration in period t. Then a smart expectation g_t^e will involve not only g_{t-1} but also earlier values $g_{t-\tau}$ for some τ larger than 1.

Just in order to reach an already familiar formula and to relate it somewhat to the discussion of this section, let us assume that the agent believes that the process of the shocks γ_t is a simple moving average of the type:

$$\gamma_t = \varepsilon_t - b\varepsilon_{t-1} = (1-b)\varepsilon_t + b(\varepsilon_t - \varepsilon_{t-1}), \tag{27}$$

where b is a fixed positive number smaller than 1 and the process of the ε_t is a pure white noise with a zero mean. Let us moreover assume that the information set J_t of the agent contains just all past values of g_t and knowledge of the stochastic nature of the shock process, as given by (27).

From his or her knowledge of all past values $g_{t-\tau}$ the agent can directly derive the corresponding values of $\gamma_{t-\tau}$ (for $\tau > 0$) and, by an iterative application of (27), the corresponding values of $\varepsilon_{t-\tau}$, which are found to be given by:

$$\varepsilon_{t-\tau} = g_{t-\tau} - (1-b)\sum_{\theta=1}^{\infty} b^{\theta-1} g_{t-\tau-\theta}. \tag{28}$$

Since $E_t\varepsilon_t = 0$, Equations (26) and (27) imply:

$$g_t^e = g_{t-1} + E_t\gamma_t = g_{t-1} - b\varepsilon_{t-1}. \tag{29}$$

Making $\tau = 1$ in (28), we find that (29) leads in this case to precisely the formula (6) applying, as we know, to the most simple form of adaptive expectations.

We note that the formation of the price level, as seen by the agent and given by (26) and (27), no longer involves his or her expectation about the future price level. The model involves no circularity. The process of the resulting rational expectations about the price level is unambiguously defined.

1.6. Empirical knowledge about the formation of expectations

For theories of inflation, and more generally for the macroeconomics of business fluctuations, correct modellizations of the processes by which expectations are formed appear to be essential. The discussion so far here suggests that there are various reasonable ideas about actual processes and that the choice between those ideas, and between the various formulas expressing them, should be enlightened by econometric investigations. There was indeed a number of such investigations to which we shall now turn attention[9].

Existing data bases are not as rich as we should like. Still today official statistical agencies tend to regard expectations as falling outside the domain of the objective facts they have the mission to regularly observe and record. Most data gathered and published on the expectations held by firms, by households or by experts come from other sources, some of which fortunately originated in the middle of this century and are still alive. A difficulty is due to the many dimensions of the phenomenon: expectations by which group of agents? On which variables? For which future date? Since expectations often are fuzzy, some sources do not collect even precise figures but rather qualitative assessments[10]. The main bulk of the material available to econometricians concerns short-term point forecasts (up to a year) on the best known macroeconomic aggregates[11].

Many possibilities are open to the econometrician approaching the subject, as to the choice of the questions to be answered and of the model within which he or she will proceed to tests or estimations. Already in the most simple case in which a time series of some expectations, for a given variable to a given horizon is available, together with the time series of the observed realization of the variable, will the econometrician decide from the start to deal with the two series as emanating from a single stable two-dimensional stochastic process? Another econometrician who would like to explore the hypothesis of rational expectations faces the question of deciding what to assume about the series of the information sets J_t: which observed variables do they contain? And which prior ideas about the phenomenon? Again, are those elements coming from a

[9] The author of this book cannot pretend to extensively know the literature on this subject, so that the survey given here is certainly not complete.

[10] Indeed, it was observed by G. Katona, an expert of surveys on individuals' intentions and expectations that most people can be induced to make a guess as to the direction of change in the near future of major macroeconomic variables, but are reluctant to give quantitative estimates of the extend of the change (G. Katona, *Psychological Economics* (Elsevier, New York, 1975).

[11] For a rare case of analysis of probability forecasts, see V. Zarnowitz and L. Lambros, Consensus and uncertainty in economic prediction, *Journal of Political Economy* (June 1987).

stable random process, perhaps subject to observable exogenous shifts? Are announced policies known? Are they taken at face value, as if perfectly credible? And so on. The wide variety of questions to be answered recommends that they be looked at in order, starting from those closest to data analysis viewed as descriptive statistics, then progressively moving to more and more elaborate inferences.

Close to data analysis is the study of errors in price expectations. Do these errors exhibit a systematic pattern? Are there particularly important ones, which would be difficult to explain? It is natural to ask how these errors correlate with the past evolution of prices. Rational expectations would imply a zero correlation coefficient if the sets J_t would contain just the data about the past evolution of the rate of inflation and knowledge of the stationarity of the process ruling this evolution. A number of studies were focused on such questions, with results that did not fit so well with the notion of rational expectations, so interpreted. However, when the information sets was assumed to also contain other data, such as series of the public deficit, the public debt, or the growth of the money supply, the fit to rational expectations was often improved[12].

A parallel line of attack searches for the estimation of good fits explaining the time series of expectations themselves. How well do fits on extrapolative or adaptive hypotheses account for recorded expectations? Which additional explanatory variables would improve the fits? This resulted in formulas analogous to those of Section 1.4, but with more lagged values than, for instance, in (7) and often also with additional explanatory variables such as the rate of increase in prices of raw materials and energy, or an indicator of the strictness of government price policy, or still a variable measuring the pressure of demand[13].

For example, P. Artus[14] analysed a quarterly series on the expectations reported by French manufacturing firms on their sales prices. Past increases in

[12] See, for instance, B. Brown and J. Maital, What do economists know? An empirical analysis of experts' expectations, *Econometrica* (March 1981); L. Jonung and D. Laidler, Are perceptions of inflation rational? Some evidence from Sweden, *American Economic Review* (December 1988).

[13] The first studies published seem to have been those concerned with analysing the answers obtained in the US from a regular survey addressed to firms' economists who were asked to give their forecasts for the next six, and the next twelve months. See S. Turnovsky, Empirical evidence of the formation of price expectations, *Journal of the American Statistical Association* (December 1970); R. Jacobs and R. Jones, Price expectations in the U.S., 1947–1975, *American Economic Review* (March 1980). For English data see J. Carlson and M. Parkin, Inflation expectations, *Econometrica* (June 1975); H. Pesaran, Formation of inflation expectations in British manufacturing industries, *Economic Journal* (December 1985). Results of the same kind were obtained from analysis of qualitative answers given by households, in G. de Menil and S. Bhalla, Direct measurement of popular price expectations, *American Economic Review* (March 1975).

[14] P. Artus, Anticipations de prix à court terme, *Annales de l'INSEE* (April–September 1977).

industrial producer prices played a predominant part in the explanation, with a term of the same sort as the right-hand member in Equation (4), coefficients b_τ decreasing by about a quarter per quarter. Almost all other identified factors concern the firms' environment: a dummy variable takes into account the price freeze in November 1963; another introduces a sharp rise in expectations after the students' revolt in June 1968; the observed rate of increase in the price of imported materials also appears. A last factor, inversely related to the pressure of demand, is the firms' available margin below productive capacity.

Given such econometric results as were available in the middle 1980s, some research was devoted to deeper investigations intended to reveal other features of price expectations. De Leeuw and Mc Kelvey showed that an aggregation bias was present in the reported data on expectations of American business firms on their sales prices and that better fits, on extrapolative formulas, were obtained at the disaggregated level[15]. Caskey gave evidence that, between 1958 and 1983, American professional economists progressively learned how to incorporate in their forecasts the autoregressive nature of the inflation process, as well as its supposed dependence on changes in the money supply[16]. Lovell attended to make a synthesis of available results on the formation of expectations by various groups of agents and drew the unpleasant conclusion that no simple specification was able to well account for the facts, whether it was inspired by the rational expectations hypothesis or by the idea of adaptive expectations; theoreticians were then advised to look at the robustness of their conclusions with respect to what they had assumed about expectations formation[17].

This conclusion is certainly disturbing and shows the limits in macroeconomics of approaches based only on prior ideas. However, in spelling out the theoretical implications of econometric results, we see that they are not all equally damaging. No real problem is presented by the fact that, in explaining expectations about the price level, some other well identified factors play a role apart from past inflation rates. If our starting point is to be, for example, Equations (2) and (4), it is not a serious problem to find, from an econometric analysis, that various events, such as a price freeze, a sharp increase in the price of imports or an exceptional increase in wages, influence expectations. These events naturally have to be included in the calculation of the exogenous component z_t appearing in Equation (2); it is then natural to take no account of

[15] F. de Leeuw and M. Mc Kelvey, Price expectations of business firms: bias in the short and the long run, *American Economic Review* (March 1984).

[16] J. Caskey, Modeling the formation of expectations: a Bayesian approach, *American Economic Review* (November 1985).

[17] M. Lovell, Tests of the rational expectations hypothesis, *American Economic Review* (March 1986).

their role when we examine how is determined what we called in Section 1.2 "the basic instantaneous trend" (even though we think that the formation of expectations is the main explanatory factor of this trend). The same consideration applies to the role of the pressure of demand on expectations, since this pressure is taken into account by the last term of Equation (2).

Before leaving the subject of empirical knowledge of expectations formation we must still take some distance from the econometric analysis of "normal times" and realize that economic history identifies exceptional cases, which do not fit into this analysis. It sometimes happened that people's expectations were completely revised after a marked change in the socio-political context. Ends of hyperinflations, like in Germany in 1948 or in Israel in 1985, provide examples of such cases. Less extreme were other cases in which government policy succeeded in provoking sharp changes in agents' expectations. The stabilizations of the French economy, attributed to Poincaré in 1927 and to Pinay in 1952, produced in people's opinion abrupt changes as regards the economic outlook, especially where inflation was concerned. We shall ignore such events in the following theoretical exploration[18], but in applied macroeconomics they may, of course, have to be recognized.

1.7. Interest rates during inflation

1. Inflation obviously perturbs financial transactions since most financial assets are denominated in nominal terms and the purchasing power of money varies. This might have been foreseen at the time these assets were created, and could have been taken into account by indexation on the price level in the specification of rates of return; but this is seldom the case. Moreover a particular attention must be brought to money, which yields a nil nominal return[19].

The perturbing effect of inflation appears particularly clear when we compare cash balances with those physical assets whose real yield would be constant and independent of the rate of inflation (assuming here this is a sufficient

[18] The study of these events is obviously interesting in itself; but it goes beyond the framework of macroeconomic theory because it has to include socio-psychological factors. Nevertheless, it would be a mistake to believe that a purely economic analysis would thus be useless or even simply secondary. In fact, history shows many cases in which socio-psychological breaks had no measurable influence on economic evolution. The peculiarity of the stabilizations we are referring to lies precisely in that the economic context associated with such breaks ensured their success; macroeconomic analysis should allow us to understand conditions for this context to be favourable.

[19] Here, and in what follows, we shall ignore time deposits and other interest bearing deposits, even though most of them are included in what is now a common definition of the money supply (M_2).

approximation for our argument). The difference in the expected yield of cash balances and that of real wealth then varies by an amount which is just equal to changes in the expected rate of inflation (assuming risk premia to remain constant). When this rate changes substantially, large shifts in the composition of asset holdings occur, and this particularly disturbs the management of liquid assets in households, firms and financial institutions. The tensions which build up as a result have an effect on non-monetary financial assets. So interest rates vary. We must study these variations in order to get a good understanding of the conditions and effects of financial policy.

It is clear, right from the start, that there is no strict and direct relationship between each interest rate and the inflation rate, since these interest rates are also influenced by monetary policy. We saw this in the previous chapter when we studied static theories where expectations were assumed exogenous, particularly with regard to future inflation. We chose one particular rate to represent all interest rates; we then considered sometimes that it was directly fixed by monetary policy, sometimes that it adjusted so as to achieve equilibrium both on the money market and on the goods market, and that it did so as a function of the money supply, then taken as exogenous. So, we cannot now pretend that the interest rate reacts purely passively and automatically to changes in the inflation rate. This particularly applies to short-term rates. But for long-term rates other reasons for change, besides those coming from inflation, also exist as we shall see later on in this section.

However, when the rate of inflation much varies, its changes may dominate all other influences on interest rates. In this case we may expect to obtain a good approximation of observed time series by applying the very simple theory according to which the real interest rate would be constant, a theory often attributed to Irving Fisher[20] who discussed it, but also showed that it did not seem to quite fit the facts.

Efforts were made to verify this theory directly, using series on expectations as to the rate of inflation. The best known is apparently Gibon's[21], who fitted the following equation for the nominal interest rate:

$$r_t = a g_{t-1}^e + b, \tag{30}$$

where a and b are two constants to be estimated. When using for g_{t-1}^e forecasts of US economists (not a representative sample of US households) for the

[20] I. Fisher, *The Theory of Interest* (Macmillan, London, 1930).

[21] W. Gibson, Interest rates and inflationary expectations, *American Economic Review* (December 1972); see also, D. Pyle, Observed price expectations and interest rates, *Review of Economics and Statistics* (August 1972).

increase in retail prices over the next twelve months, he found a value of a very close to 1 in the equation for the interest on loans over the same period. The fitted equation for the rate on ten year loans gave a smaller value for the coefficient a $(3/4)$, which could be the proof that the idea of a "return to normal" plays a part in long-term expectations (the idea indeed implies that these expectations are less variable than those over twelve months, used in the fit).

We note that, if Equation (30) holds and if price expectations are extrapolative, past rates of inflation are determinants of present nominal interest rates. Inversely, if the same equation holds and if price expectations are good predictors of future rates of inflation, changes in nominal interest rates precede and signal future changes in the inflation rate, a phenomenon sometimes identified with "Granger causality" from interest rates to inflation rates. The dynamic pattern involving such mutual lags was discussed in various contexts and is worth remembering[22].

2. Be that as it may, we shall here keep the definition of the expected real interest rate ρ_t^e that was given in Chapter 1, Section 2.5, and that follows from equation[23]:

$$r_t = g_{t-1}^e + \rho_t^e \tag{31}$$

but we shall not always assume that ρ_t^e is constant. On the contrary the real rate will often be assumed to depend on monetary policy which determines, more or less directly, the short-term nominal interest rate. In some other cases it will be assumed to depend on shifts in demand and supply on the market for loans.

A brief examination of econometric studies on fluctuations of interest rates is useful at this stage in order to provide some knowledge of the context for

[22] See, for instance, M. Gallager, The inverted Fisher hypothesis: additional evidence, *American Economic Review* (March 1986); R. Barsky, G. Mankiw, J. Miron and D. Weil, The worldwide change in the behavior of interest rates and prices in 1914, *European Economic Review* (June 1988).

[23] The reader may note a slight difference with Chapter 1, in the fact that we now write in the formula g_{t-1}^e instead of what would now correspond to g_t^e. This change, which does not much affect the following results, comes from a different assumption on the evolution of the price level. We had explicitly assumed at the beginning of Section 1.3 of Chapter 1 that this level p_t would remain constant from date t up to the end of period t and jump at date $t+1$ to the new value p_{t+1}. In better conformity to the facts we are now taking into account the evolution within the period: p_t is the average price level during the period and g_t its rate of increase to the average during the next period. At the time of a loan contract running over period t and concluded at date t with the rate of interest r_t, the rate of increase in prices between period $t-1$ and period t provides the most natural reference. It is not yet known at date t and must be anticipated, hence the formula.

which the models in the following parts will be exposed. In addition, this may help to bring here new evidence as to the expressions which are likely to best represent price expectations. The econometric studies to be considered concern two distinct questions. How does a representative rate change over time in relation to various other macroeconomic variables? How do the various rates of interest change in relation to each other, especially interest rates over different maturities? The first question interests us most; we shall speak of the second one still more briefly.

In neither of the two cases are conclusions coming from econometric results precise. This is a serious drawback for macroeconomics, given the part played by real interest rates in shaping the business climate (we should like for instance to better explain the persistence of high rates throughout the 1980s and most of the 1990s). The difficulty is linked in particular to the high and varying volatility observed on financial markets, which is itself linked to the role of expected capital gains or losses on holdings of foreign assets, on stocks of shares, or on assets invested in tradable long-term loans. Important risk premia are involved, which are likely to vary with uncertainty on future quotations and with the liquidity of the markets.

To give a proper account of the fluctuations in the representative interest rate, theory suggests that an econometric equation ought to contain three types of terms:

- the first would come from the past trend in the growth of the price level (or from expectations as to the future rate of inflation); this is the so-called "Fisher effect", our main concern here;
- the second would reflect changes over time in those factors which affect the goods market, or equivalently the conditions for the IS equilibrium; these could be saving propensities or rates of return expected from productive operations; such factors are likely to be related to the rate of growth of production, to an indicator of the stand of fiscal policy, etc.; we sometimes talk about the "Wicksell effect";
- the third would represent changes in the factors affecting the money market, or equivalently the LM equilibrium; there could appear an indicator of either the money supply, or of the conditions affecting this supply (the monetary base or something else); we then speak of the "Keynesian effect".

We shall not dwell on these names which are historically disputable, for each of the three famous economists was aware of all three effects. The main purpose is here to suggest a few references to the reader, where the roles of the above factors have been estimated.

The first well known references to detailed estimations seem to be those concerning long-term rates in the United States; one is based on annual data from 1902 to 1940, the other on quarterly data from 1954 to 1969: T. Sargent, Commodity price expectations and the interest rate, *Quarterly Journal of Economics* (February 1969); M. Feldstein and O. Eckstein, The fundamental determinants of the interest rate, *Review of Economics and Statistics* (November 1970). For a later reference, which uses in particular the forecasts for inflation made by economists in US firms and which concerns the period 1952–1975, see also V. Tanzi, Inflation expectations, economic activity, taxes and interest rates, *American Economic Review* (March 1980).

For the more recent literature a guiding issue concerns the main reasons given of the high levels of real interest rates in the 1980s. A frequent explanation has been the impact of public deficits on the saving-investment equilibrium[24]. But two other lines of explanation were explored by Blanchard and Summers[25]: the high levels of interest rates may be due to the expectation of high profit rates on productive operations and was indeed accompanied with increases in prices of shares on the stock exchanges; it may also be due to portfolio restructuring because of changes in risk premia and in covariances between various asset returns, as well as between those returns and the flow of consumption. Looking at the time series of seven large OECD economies for the period 1959–1988, Barro and Sala-i-Martin[26] concluded that improvement in profit prospects played a major role, and monetary policy a significant one, but that public deficits did not seem to have mattered (their fit gives also a surprisingly important role to changes in the price of oil).

Going back to our starting theme, the impact of changes in the inflation rate on the interest rate, we must note that the real interest rate, usually measured as the difference between the nominal rate and the contemporary rate of inflation, appears to be particularly low when prices increase quickly (we earlier related this observation to the idea that price expectations might be regressive). The phenomenon was attributed to money illusion first by I. Fisher himself, more recently by Modigliani and Cohn; it was further studied by L. Summers[27].

[24] M. Feldstein, US budget deficits and the European economies: resolving the political economy puzzle, *American Economic Review* (May 1986).

[25] O. Blanchard and L. Summers, Perspectives on high world real interest rates, *Brookings Papers on Economic Activity*, No. 2 (1984).

[26] R. Barro and X. Sala-i-Martin, World real interest rates, in: O. Blanchard and S. Fisher (eds.), *NBER Macroeconomics Annual* (MIT Press, Cambridge, 1990).

[27] R. Thaler, Irving Fisher: modern behavioral economist, *American Economic Review* (May 1997); F. Modigliani and R. Cohn, Inflation, rational valuation and the market, *Financial Analysts Journal* (1979); L. Summers, The nonadjustment of nominal interest rates: a study of the

There seems to be agreement now that, in order to maintain the "Fisher hypothesis" as given by Equation (30) with $a = 1$, one has to assume that a very long lag is required for price expectations to adapt to the actual movement of prices[28].

3. Let us now turn to a brief survey of what is known about the explanation of changes in the term structures of interest rates, i.e. changes in the "yield curve" plotting, at a given date, the interest rate of bonds of different maturities as function of their respective terms (a day, a quarter, a year, ten years, ...). According to B. Malkiel (in Eatwell et al., *New Palgrave*, op. cit.), historically three competing theories have attracted the widest attention. According to the "expectations theory" the shape of the yield curve can be explained by investors' expectations about future interest rates, because of the possibility of arbitrage: an investor who needs to recover his or her funds in a year can consider either issuing bonds for a year or issuing bonds for three months now, but already decide to issue again such bonds one, two and three quarters later; the no-arbitrage condition gives an equation between the rates for term of a year and for term of a quarter, together with expected rates for the later three month issues.

The "liquidity-preference theory" accepts such equations as an initial reference but argues that, in a world of uncertainty, short-term issues are more desirable to most investors than longer-term issues because the former are more liquid: the price at which they will be sold if and when the funds will be needed is less uncertain. Long-term bonds have to yield an additional "liquidity premium".

The "preferred habitat theory" argues that the case just considered is too much focused on the situation of liquidity managers and that other investors, like life insurance companies, prefer long-term issues in order to hedge against the risk of interest-rate fluctuations. If all investors were fully hedging, each one motivated by his or her future needs for funds and no one speculating, short and long rates would be fairly independently determined by supplies and demands of corresponding bonds on their respective markets.

Nothing prevents in principle to merge all the above considerations within a model of equilibrium on financial markets, but a lot of information would be needed for the model to deliver useful predictions to macroeconomists. The latter are therefore interested by econometric results exhibiting the main regularities which are likely to explain changes in the yield curve.

Fisher effect, in: J. Tobin (ed.), *A Symposium in Honor of Arthur Okun* (The Brookings Institution, Washington, 1983).

[28] See M. Viren, Examining the long-run relationship between interest rates and inflation: some cross-country evidence, *Economic Notes*, by Monte dei Paschi di Siena, No. 3 (1986).

From his survey of empirical studies, Malkiel concludes that expectations do influence the shape of the yield curve, but that the pure form of the expectations theory has generally be rejected. There does appear to be an upward bias to the shape of the curve, indicating that term premia exist. But, contrary to the liquidity-preference theory, these premia vary over time and do not increase monotonically over the whole span of forward rates.

1.8. Asset prices

Since the implementation of national accounting and the advent of applied macroeconomics fifty years ago, the common practice is to measure the pace of inflation by the growth rate of an indicator concerning the flows of either productions or consumptions. Most often the indicator is the GDP deflator or a consumer price index, the latter being available monthly. But we should not forget that, at times, major price changes may concern stocks, that is assets, rather than flows. Important ups and downs in prices of real estates or equities recurrently occur, often ending in financial crises. We shall not pay much attention to the phenomenon in this chapter, because it played little role in the mainstream development of macroeconomic thinking about inflation. As we shall see in Chapter 9, Section 2.10, the phenomenon played a part in some theories of the business cycle. Here is the proper place for a brief examination[29].

1. All assets for which the notion of a market price is meaningful may experience large movements. This applies to farm land, forests or urban land, to residential as well as commercial real estates, to equities of enterprises quoted on stock exchanges, to national currencies on international markets, to bonds in relation with movements of long-term interest rates, to works of art, to famous wines, ...; it even applied to tulip bulbs in the seventeenth century. Such movements are, of course, very significant for some private businesses and for the wealth of some families. What is their macroeconomic impact?

Some argue that it is close to be negligible, unless monetary authorities badly react to them. For instance, M. Friedman and A. Schwartz, in their famous work to be quoted later in this chapter, commented that monetary authorities should have ignored the rise in the New York stock market in 1928 and 1929, and rather focused attention on other goals such as the general price

[29] A good account, suggesting the importance of the phenomenon for macroeconomists, is given by C. Kindleberger, Asset inflation and monetary policy, *Banca Nazionale del Lavoro Quarterly Review* (March 1995).

level. Similarly the US administration openly claimed to practice "benign neglect" when the dollar was thought to be overvalued in the early 1980s.

A relevant question in this respect is to know whether asset prices are leading indicators of changes in inflation as commonly measured, or leading indicators of output and employment. C. Kindleberger (op. cit.) quotes some interesting work done in this respect at the Bank for International Settlements, where indexes of aggregate asset prices are now computed[30]. The evidence shows that asset prices cannot be taken as reliable leading indicators of inflation. It is not so much that asset prices are very volatile, but there are cases of divergent evolutions between asset prices and the current measures of the inflation rate. Although stock exchange indices of equity prices belong to the set of indicators used by business analysts, it is not fully reliable for variations in output either: the 1987 crash was actually followed by the extension of the boom to 1988.

The volatility of asset prices can be rationalized by the volatility of their expected rates of return. During each period this return is made of two parts, as we already saw in Chapter 1 (Sections 2.6 and 2.7): the income of the period and the capital gain. Anything that leads to revise the expectation of the future prices of an asset means also a revision of its expected returns because of a new assessment of capital gains or losses; the latter revision results in an immediate change in the price of the asset. The dependence of the current price on the prices expected to hold in the future, sometimes a very random future, naturally leads to volatility.

By contrast prices of output or consumption are closely dependent on production costs, and this may explain a divergent evolution. In particular a cost-push inflation is often due to large wage increases, which are likely to hurt profits, the income part of the return on equities. If so, the price of equities simultaneously falls.

2. A debated issue is to know whether the volatility of asset prices can be fully explained by the stochastic properties of the expected flows of incomes earned by the assets, or whether it also comes from the existence of "bubbles". We speak of a bubble on an asset when we think that the price is high (low) today only because investors believe the selling price will be high (low) tomorrow, whereas the "fundamental" factors of the incomes to be earned on the asset do not seem to justify such a price[31]. We are then hinting that the

[30] See in particular C. Borio, N. Kennedy and S. Prowse, Exploring aggregate asset price fluctuations across countries: measurements, determinants and monetary policy implications, *BIS Economic Papers*, No. 40 (Bank for International Settlements, Basle, 1994).

[31] See J. Stiglitz, Symposium on bubbles, *Journal of Economic Perspectives* (Spring 1990).

situation will not go on for ever on such a basis and that someday the bubble will burst. Saying that bubbles exist so amounts to stating that asset markets are not perfectly ruled by the simple rational calculations taught in courses on microeconomic theory.

Various approaches are followed in attempts to settle the debate. The first one remains in the spirit of microeconomic theory and ask such questions as: in the absence of a complete set of futures markets, extending infinitely far into the future, would the sequence of competitive temporary equilibria rule out any path with a bubble?[32] Or, could not bubbles persist for a long time in markets where agents have limited lifetimes and limited access to capital?

The second approach looks for econometric tests that, using time series data on the prices and incomes of assets, could reject the hypothesis of no bubble. But this approach faces the problem of separating the apparition of bubble movements from the possibility of misspecification of the model giving the underlying stochastic process of incomes accruing to the asset (the process of the fundamentals)[33]. Whether or not this possibility is put forward beyond reason against tests rejecting the hypothesis of no bubble, is a question left to the appreciation of readers of the literature.

If many economists believe in the occasional occurrence of bubbles it is because their approach is to look at the perception of market participants in cases when strong ups and downs of prices were observed: stock market crashes were then commonly attributed to the idea that the rates had earlier increased too much, not to the idea that recent news had led to a sharp downward revision of future expected incomes; similarly, real estate dealers suffering from a depression of real estate prices often referred back to what had happened a few years earlier when prices were booming and investors were subject to a shortage illusion. R. Shiller organized surveys in order to better follow what market participants were indeed saying[34]. After the 1987 stock market crash the most common answer was that the market had been overpriced, an idea related to "a theory about investor psychology"; only a rather small minority of respondents related their explanation to "a theory about fundamentals such as profits or interest rates". The reported incidence of investor psychology was much lower in booming or depressed real estate markets; but respondents did not refer to quantitative evidence about future trends in supply or demand, or the professional forecasts of future supply or demand; purely speculative

[32] The question is then close to the ones discussed in Section 5.4 of our Chapter 6.

[33] R. Flood and R. Hodrick, On testing for speculative bubbles, *Journal of Economic Perspectives* (Spring 1990).

[34] R. Shiller, Speculative prices and popular models, *Journal of Economic Perspectives* (Spring 1990).

and simple-minded considerations were found to be a prime motive for buying homes in boom cities (such as: "unless I buy now, I won't be able to afford a home later"). R. Shiller concluded that market participants refer to "popular models" which are often unsophisticated, nevertheless complex to describe, but certainly different from those used by theorists, at least in dramatic boom and burst situations.

3. Notwithstanding what was reported earlier in this section, it is hard to believe that large movements of asset prices have no real macroeconomic effects. Businessmen recurrently complain about the overvaluation of their home currency, which hurt their exports, or about the unpredictability of exchange rates, which may quickly jeopardize the profitability of their investment in building networks abroad. Many financial institutions underwent large losses after stock-market crashes, or after unexpected drops in bond prices, or still after depressions in real estate prices. Consequences of these financial losses were often magnified by sudden needs for liquidities, which could only be satisfied by the sale of assets at quite unfavourable conditions. Monetary authorities could not help feeling challenged at such times, and indeed they sometimes deliberately injected money into the economy, like after the 1987 crash.

C. Kindleberger (op. cit.) considers a different question namely not whether volatility of asset prices has real effects but whether monetary authorities should intervene in anticipation in order to dampen "asset inflations" when they become important. He reports on a number of cases when the problem was discussed since the interwar period. It is a problem because we find it difficult to work out a rule to restrain asset inflation without running the risk of harming output and employment. Central bankers cannot earmark liquidities for current productive operations; at times of asset inflation the money supply is used as well for fuelling speculative buying of assets; restraining the money supply would contribute to dampen asset inflation but would also hurt the normal financing of business. One instrument cannot serve to reach two goals when they would require opposite actions. The difficulty must be kept in mind because we shall not mention it again in this chapter, where discussion of monetary policy will, however, play a large part.

1.9. The costs of inflation

It is clear that economic policy must have, among its objectives, the fight against inflation, as well as the fight against unemployment. Inflation and unemployment seem to the public at large to be two illnesses which affect the functioning of the economic organism, two illnesses which the government has the duty to remedy[35].

Nevertheless, the economist who pauses in order to think about the reasons for protests against inflation realizes that this is not a simple question. The sentence, "if inflation continues, it is because everyone gains from it" is certainly a paradox[36]; but the sentence would not be so often heard if rejecting it out of hand was easy. The theme of the costs of inflation has thus found its place in the economic literature, and we should examine it here. We realize, however, that a purely economic analysis may be insufficient, since the costs of inflation are not all of an economic nature (neither are all costs of unemployment). Some of the "social" costs of inflation would be better studied by sociologists or lawyers.

Whenever one thinks precisely about the costs of inflation, one is led to distinguish between two cases depending upon whether the price increase was anticipated or not. The effects, and thus the costs, are different in the two cases. But the distinction is not clear cut. For the theorist, the temptation is to deal, first and foremost, with the pure case where inflation would be perfectly anticipated and to defer consideration of less well defined cases. Unfortunately, this pure case is too extreme and far from reality to be a reasonable reference point.

Indeed, in the pure case not only did some particularly perceptive individuals correctly anticipate inflation, but also all agents did so and all took this perfect expectation into account in all their dealings with each other. It then becomes natural to assume that institutions themselves have adapted to the situation, so that those who have not should be clearly identified. One will not go as far as assuming that a new and stable money was introduced and is widely used as a means of payment (this would however be the complete institutional response), for inflation would have really disappeared with respect to this new money. But one could go as far as assuming that accounting norms, commercial and conventional practices, the tax system have all perfectly adapted, an assumption which is inconsistent with reality. In other words, the pure case can only be used when dealing with certain special costs, where it does not seem

[35] It was often noted that people dislike inflation more than economists seem to do. See in particular R. Shiller, Why do people dislike inflation?, NBER Working paper No. 5539 (April 1996).

[36] Obviously, it would be very different to say that everyone profits and suffers from inflation. Perhaps the quoted sentence should be read as just referring to the fact that, in the inflationary spiral where increases in prices and wages follow each other, each person profits from some of these increases and no one is ready to forego those which are in his or her own interest. The sentence so understood does not get us very far and it has the shortcoming of supporting the view that inflation might be harmless.

to strongly bias the results of the analysis. This severely restricts its domain of relevance.

Inflation is the increase in the overall price level, or, what amounts to the same thing, the depreciation of the money. We have to know how it perturbs the working of each of the three functions of money: means of payment, store of value, nominal reference for all prices.

Inflation has to be very acute in order for money to be no longer used, literally speaking, as a means of payment. But any inflation increases transaction costs, whereas the use of money has precisely the purpose of reducing these costs to a minimum. Indeed, inflation results in a decrease in the purchasing power of cash balances and in an increase in nominal interest rates, hence in an increase in the opportunity cost of holding money rather than bonds; economic agents are thus led to hold less cash; they then have to accept higher transaction costs than they would have otherwise in some of their operations (more trips to the bank, smaller reserves in easily convertible financial assets). This is the case, because money, although still a means of payment, has become a worse store of value.

It is possible to provide an estimate for this cost of inflation, starting from an empirical equation of the demand for money and knowing in particular its elasticity with respect to the interest rate. The results may be a little biased downward in that they do not take proper account of the substitutability between the holding of cash balances and the use of assets for consumption or investment. The qualitative conclusion is, nevertheless, worth recording: the decrease in liquidity which is produced by perfectly anticipated inflation would have a very low cost[37]. It has been argued that such estimates ignore the fact that banking practices do not adapt to inflation: the amount of credits granted to customers remains fixed in nominal terms; the increase in nominal interest rates means higher "front end loading" in the service of loans; overall, liquidity problems are increased by an effectively tighter credit rationing.

Inflation costs may arise more from the deterioration of the money unit as a standard reference for all prices. A monetary payment appears as a counterpart in almost all transactions; in many contracts it is impossible to precisely date the transaction or its counterpart. Hence, inflation introduces disorder into economic relationships, because the standard unit of value depreciates with time. The symptoms of this disorder, and all its consequences, are too numerous and varied for it to be possible to describe and evaluate them all. Many of them

[37] See R. Barro, Inflationary finance and the welfare cost of inflation, *Journal of Political Economy* (Sept.–Oct. 1972); D. Laidler, The welfare costs of inflation in neoclassical theory; some unsettled problems, in: E. Lundberg (ed.), *Inflation Theory and Anti-Inflation Policy* (Macmillan, London, 1977).

affect the fairness of contractual or legal relationships and give, to the cynical, the opportunity to unduly benefit from the scrupulous. This is exactly why inflation is often thought of in the same way as are loopholes in the law[38]. It produces the same sort of protests and dissatisfaction.

Some of these manifestations can be studied; we can indeed analyse how inflation reacts on the way in which prices are fixed and announced, especially in the case where this is done through catalogues. Since it is costly to change price-lists, the higher inflation the less strictly do prices follow the path which should result from their fundamental determinants[39]. Since changes made by different producers and sellers are not synchronized, there are also resultant erratic changes in relative prices.

This last result has been examined empirically, since it is easy to test whether inflation increases the spread of short-term changes in relative prices. Econometric tests are actually less conclusive than we might have thought. They suggest that we have to get right down to the fine details of product and date in order to clearly find this phenomenon and that its formulation may have to be amended. According to Frochen and Maarek[40], it is mainly phases of rapid change in the rate of inflation which produce a high level of disturbance in relative prices.

Inflation seems to be associated with other erratic trends. So, the thesis was put forward, and to some extent econometrically supported, that countries and periods in which inflation is on average the most rapid are also those in which the rate of price increase changes the most: a strong tendency for money to depreciate corresponds to a rate of depreciation which fluctuates wildly in the short run[41]. Similarly, it has been shown that the forecasts of different eco-

[38] This point was argued very forcefully in A. Leijonhufvud, Costs and consequences of inflation, in: G. Harcourt (ed.), *The Microeconomic Foundations of Macroeconomics* (Macmillan, London, 1977).

[39] E. Sheskinski and Y. Weiss, Inflation and costs of price adjustment, *Review of Economic Studies* (June 1977).

[40] P. Frochen and G. Maarek, Inflation et déformation des prix relatifs, *Cahiers économiques et monétaires*, Banque de France, No. 8 (1978). See also D. Jaffee and E. Kleiman, The welfare implications of uneven inflation, in: E. Lundlberg (ed.), *Inflation Theory and Anti-Inflation Policy* (Macmillan, London, 1977); S. Fischer, Relative shocks, relative price variability and inflation, *Brookings Papers on Economic Activity*, No. 2 (1981); A. Cukierman and P. Wachtel, Relative price variability and non-uniform inflationary expectations, *Journal of Political Economy* (February 1982); S. Domberger, Relative price variability and inflation: a disaggregated analysis, *Journal of Political Economy* (June 1987); M. Reinsdorf, New evidence on the relation between inflation and price dispersion, *American Economic Review* (June 1994).

[41] See the discussion of this thesis in D. Jaffee and E. Kleiman, op. cit., and in G. Ackley, The costs of inflation, *American Economic Review* (May 1978).

nomic agents as to the future rate of price increase were all the more dispersed as inflation was faster or more erratic[42].

It is intuitively clear that, if inflation makes later price increases more difficult to forecast, then the result will be various costs, linked to the fact that agents make less appropriate decisions, and to the fact that capital gains or looses become more random. These costs, which concern both the efficiency in the allocation of resources and the fairness in economic transactions, are obviously difficult to evaluate.

However, we must also note that a possible way of explaining the fact that inflation becomes more variable as its rate increases is to show that economic agents often succeed in avoiding part of the costs we have just been talking about. Indeed, the spread of the practice of indexing constitutes a spontaneous institutional response to sharp and erratic inflation: applied to lasting economic relationships (work contracts or long-term loans, for example), indexing aims to ensure that the purchasing power of what the contracting parties receive remains stable. But the spread of indexing is likely to result in an increase in the rate of inflation, because the depressive effect of price rises on aggregate demand will be mitigated, as we shall argue in Sections 2.4 and 4.1. Widespread indexation is, therefore, a very imperfect institutional response; it certainly reduces inflation costs for indexed economic operations, but it tends to increase costs for non-indexed operations.

Among these last operations, taxes have a particularly important place. Whilst tax scales can be revised often enough to keep up with inflation, this is not the case for the actual structure of the tax system which, like all other social "rules of the game", must be maintained over a long period to allow people to become familiar with it. This structure is thus not revised often and only after laborious bargaining. Neither taxes, nor firms' accounts on which taxes are based, are constructed to take inflation into systematic consideration. A Deutsch mark is accounted for as a mark no matter the date at which it is paid in or paid out. In times of inflation, taxes thus introduce large distortions, which are foreign to the principles on which they are based. The allocation of resources is thus disturbed.

Speaking of taxation we must recognize in particular two significant effects of inflation. The first one, often called the "Tanzi effect", comes from the fact that many taxes are collected and payed after a delay, which reduces their value in real terms. The second one, the "inflation tax", is imposed on holders of cash, whose real value depreciates; it benefits the banking sector, which issues

[42] See again D. Jaffee and E. Kleiman, op. cit., and A. Cukierman and P. Wachtel, Differential inflationary expectations and the variability of the rate of inflation, *American Economic Review* (September 1979).

money, more particularly the state-owned central bank; issue of money, which is demanded in order to replenish real cash holdings, provides free financing to the state, the “seigniorage”. The inflation tax is clearly regressive, because it is light for wealthy people; the Tanzi effect concerns mainly direct taxes and so benefits mostly the same people.

More generally, questions are often asked as to the redistributive effects of inflation: between firms, households and the government, between social categories, between the rich and the poor. As was suggested a little earlier, those among such effects that are observable at the macroeconomic level are neither the only ones nor perhaps the most serious, for the inequitable redistribution between agents of the same category may be even more important. Macroeconomic analysis of the redistribution produced by inflation is, moreover, subject to the inherent difficulty of all studies of redistribution, that is, knowing how to choose the right reference point; this reference is a fictitious situation meant to represent what the distribution would have been if redistribution had not occurred. The earlier questions of knowing to what extent inflation is expected and to what extent institutions adapt to it cannot be avoided.

Assuming here that we are interested in knowing what are the medium-term redistributive effects of an unanticipated acceleration of inflation and without wanting to push the analysis of this question too far, we can stick to the statement that this inflation results mainly in reducing the purchasing power of claims and in lightening the real burden of debts. It thus benefits debtors at the expense of creditors, that is, it profits firms and the government to the detriment of households; it also benefits young people and the middle social classes (the most in debt) to the detriment of old people, of the wealthiest social classes, and also of the poorest (who suffer from the loss of purchasing power of their cash holdings, who have hardly any debts because rare are those giving them credit, and who also suffer from a less well informed management of their resources and often from a poorer indexing of their income).

In Part 7 of this chapter we shall moreover take account of short-term redistributive effects which play a role in the dynamics of the inflationary process and concern the relationship between employees and employers.

Finally, let us remember that inflation has a high aggregate cost whenever it clearly damages economic growth. Such is the case at very high inflation rates (more than 20 per cent annually, say). But the relation between inflation and growth is less clear at moderate rates, as we argued in Chapter 5, Part 6 (see in particular the empirical references given in Section 6.2 of that chapter).

2 Pressure of Demand and Inflation

Of all the topics which concern inflation the one most discussed in economic theory concerns the reciprocal relationship between the aggregate demand for goods and services and the increase in the price level, starting with the specification of the last term in Equation (2). This is not surprising since the relationship obviously falls within the domain of economic analysis, whereas factors acting on expectations or on costs may come from other domains. Moreover, common anti-inflationary policies operate by squeezing demand; it is mainly left to economists to evaluate under what conditions such policies reach their objective.

This part, devoted to a first study of the reciprocal relationship, will be limited in two respects, concerning one the representation of the economic system, the other the examination of policies. As for the economic system, arguments will introduce only a very small number of variables. In particular, the labour market, which will play an important part later in the chapter, will not explicitly appear for the time being. But the formalizations will aim at taking proper account of the main dynamics of the phenomenon. They will lead to systems which rapidly become complex in their utilization, even when few variables are introduced. In getting first familiar with simple modellizations, simple econometric investigations and simple logical arguments, we shall prepare ourselves for the later study of specifications which will contain a larger number of variables. So, for the moment, we are not claiming to give a full account of the facts, but rather to introduce fairly simple ways of reasoning, to show the essential steps in the arguments and to study some of the conclusions reached.

Such arguments concern the short run. The dynamic medium which they introduce should be thought of as intending to be adequate for developments over one, two or three years, thus over a dozen quarters at most. Extending their solutions further may be interesting in order to understand properties of these solutions, but does not constitute a test of the models, which are not meant for long horizons. Other models are required to deal with the relationships between growth and inflation (see Chapter 5, Part 6), or even to determine the main periodicities of business cycles (Chapter 9).

Since we are not going to pursue here the medium and long-run consequences of shocks or economic policies, we can moreover take productive capacity as exogenous, which will simplify things considerably. The development of capacities will be considered as resulting from past evolution. Actually, the models, which will ignore long-run trends, will also assume for simplicity a constant productive capacity.

About economic policies, this part will just pose some preliminaries for what will be discussed systematically in the two following ones. Such a delimitation will permit a clear distinction between the two approaches to policy problems that will be the subjects of respectively Part 3 and Part 4.

The dynamic relationship between the evolutions of the demand pressure and of the price level will figure at the centre of this part. It will explain a dilemma confronting aggregate demand policy when it has to slow down inflation (Section 2.2). This relationship results from the way in which prices and quantities adjust in the short run (Section 2.1). But there is also a feedback from price changes to the level of demand, and we shall have to take this into account, especially when we shall study the impact of monetary policy on inflation. This is why we must first consider how consumers and investors are likely to adjust their demands in the short run (Section 2.4). Even before that, we shall (in Section 2.3) want to get a feeling of the empirical properties of the dynamic relationship between prices and quantities at the most aggregate level.

2.1. Short-run adjustments

In the same way as in Chapter 6 we may here refer to the vision of a moving temporary equilibrium. At any given time the state of the economy determines, more or less closely, how the economy will further evolve. To fully characterize the state and its evolution, we would need to give the values taken on by a whole set of very many variables. But, particularly in this chapter and at this point, we boldly simplify.

At the most aggregate level to which we now stand, and considering only the quantities and prices of goods and services, we need very few variables to characterize the state of the economy. Apart from the volume of output y_t and the price level p_t, we include productive capacity $\bar{y}$, expectations as to the rate of price increase g_t^e and the volume of aggregate demand D_t. We shall come back later, at various places, to the definition of productive capacity, already in Section 2.3 below. At this stage we may simply refer loosely to the concept, as we already did in Section 3.1 of Chapter 7.

But the definition of aggregate demand deserves a mention because it will not exactly correspond to the one chosen for the aggregate static models in Chapter 7. The inequality $y_t \leqslant D_t$, which figured in these models, will no longer be taken as necessarily holding. When studying short-run dynamics it is indeed interesting to allow for gaps, either positive or negative, between "demand" and production. Aggregate demand will not be a purely *ex ante* notion; it could even be measured in terms of observed flows. In comparison with national accounting practices, it would then be a little unusual. It would not cover the "involuntary" accumulation or depletion of inventories, which suppliers have to accept because their sales do not exactly correspond to their intentions. It might include, on the other hand, standing orders waiting to be delivered to buyers, particularly when sellers find they have too many orders to fill, and thus have to ration their customers to some extent. To sum up, this variable may be viewed as lying somewhere between that found in national accounts and the purely conceptual notion of the short-term static theory of Chapter 7.

Two additional variables, simply derived from the volumes y_t, $\bar{y}$ and D_t, will play important roles in the dynamics of inflation, two "indicators of tension" as it were: the degree q_t of utilization of productive capacity and the intensity d_t of demand pressure:

$$q_t = \frac{y_t}{\bar{y}}, \qquad d_t = \frac{D_t - y_t}{\bar{y}}. \tag{32}$$

A temporary short-term equilibrium, meant to describe the instantaneous state of the economy, will be defined at the same time as the adjustment laws that we shall consider presently. The determination of this equilibrium will not be explicitly linked to that of the equilibria discussed in the previous chapter. Nevertheless, as we shall see, it will fundamentally be an equilibrium of the same sort. The adjustment process will specify how the four variables p, q, g_p^e and D evolve. Evolution of D will result from the behaviour of individuals and from the choices of economic policy; it will be characterized in Section 2.4.

For simplicity evolution of g_p^e will most often be assumed to be ruled by Equation (6) or a simple generalization of it; we have already discussed how this might be justified.

We must now characterize how p and q simultaneously evolve. As in Chapter 7, Section 2.1, we have to formalize current ideas as to short-run adjustments in prices and productions. We must not only include expectations as to future prices, but also date the reactions. We have moreover to favour expressions which will be simple enough to make subsequent arguments easy (this already prompted us to refer to Equation (6) rather than to the more general equation (4) for the formation of expectations).

The following ideas will now be formalized:

- if the degree of capacity utilization is normal, production increases or decreases depending upon whether the pressure of demand is positive or negative; if not, production also tends to change in the direction of a return to normal capacity utilization;
- the higher the pressure of demand and the degree of capacity utilization, the higher the speed of price increase.

In Part 8 of this chapter we shall examine some of the theoretical explanations which can be given for reactions of this type. We shall then consider in particular adjustment costs, as well as uncertainty implying precaution when assuming these costs. For the time being we just state that the above ideas are based on observation.

More precisely we shall characterize short-run adjustments by the two equations:

$$q_{t+1} = q_t + \varphi(d_t, q_t), \tag{33}$$

$$g_t = g_t^e + \pi(d_t, q_t), \tag{34}$$

where the two functions φ and π are increasing with the pressure of demand d_t, the function φ decreasing in q_t, whereas function π is increasing with this variable. Moreover, the two functions φ and π take the zero value when $d_t = 0$ and $q_t = \hat{q}$, where $\hat{q}$ represents the "normal" degree of capacity utilization. So are the ideas expressed above taken into account.

Remember that to determine g_t is an fact determining p_{t+1}, as Equation (1) shows; the two equations (33) and (34) therefore determine the volume of production and the price level of period $t+1$ from values observed in period t for demand, production, the price level and its recent evolution. For the assumption so made to be tenable, the unit period must be fairly short. We shall fix this in our minds by thinking of a quarter, the sort of length which seems admissible for all models in this chapter. The lag in the reaction of y_t and p_t to D_t clearly constitutes a significant difference with the adjustments specified in Chapter 7, Section 2.1. The static Chapter 7 could not take this lag into account. But remember that the short-term equilibrium of Chapter 7 was conceived as having to apply to the analysis of a longer period than a quarter: a year, say.

Consequently there is consistency between the adjustment models chosen in the two chapters if the pressure of demand and capacity utilization remain constant for a certain length of time, a condition which, taking (33) into account, is the same as:

$$\varphi\left(\frac{D_t - y_t}{\bar{y}}, \frac{y_t}{\bar{y}}\right) = 0. \tag{35}$$

Contrary to what was assumed in Part 2 of Chapter 7, there is not always equality between effective demand and production, even with a stationary temporary equilibrium; but this comes from a different definition of "aggregate demand". There is strict equality $y_t = D_t$ when there is normal use of capacities ($y_t = \hat{q}\bar{y}$); if not, there is a discrepancy: production stays lower than demand when capacity utilization is abnormally high; it is higher in the opposite case.

Leaving aside this difference, Equations (32) and (34) present a view of adjustments which is close to that illustrated by Figure 1 in Chapter 7. Continued high pressure of demand over several successive quarters tends to lead to an acceleration in the rate of inflation if the use of productive capacities is high; adjustment is thus made through prices rather than through quantities (point P in Figure 1 below). The same pressure of demand leads, on the contrary, to an increase in production if use of capacities is low (point P'). The nonlinearity of adjustments in Chapter 7 is also found here and low pressure of demand over a long period produces, in the case of high use of capacities, an adjustment through quantities rather than through prices (point P'').

Taking here into account adjustment lags in output and the price level, we clearly find situations close to those studied by means of the temporary equilibria of the previous chapter, especially those that were the object of its Part 2 where, as in the present part, the labour market was not explicitly introduced.

We should still make two quick remarks about the system of adjustments defined by (33) and (34). On the one hand, it is a little more elaborate than Equation (2) set out initially in this chapter to characterize the inflationary process, since it explicitly includes the degree of capacity utilization. On the other hand, it gives a minimal room to expectations, since only those concerning prices are taken into account, and in a way that excludes any shock breaking them from the past. Clearly, there may be sharp changes in expectations as to demand; then productions or prices have a good chance to be revised as a result, even before the new pressure of demand is being felt. The same sort of

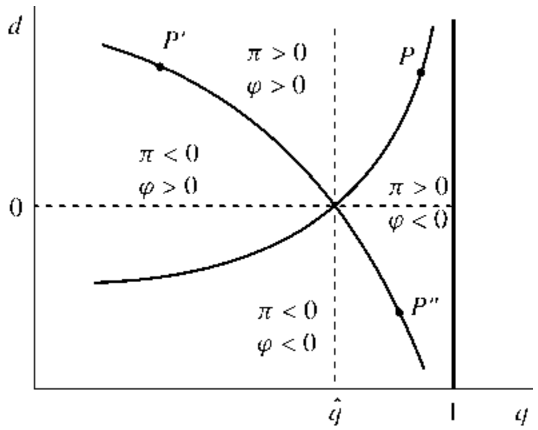

Fig. 1.

early reaction is likely if there is a sharp change in expectations as to the future rate of inflation. This remark concerns in particular the observed features we shall discuss later when shocks will have to be introduced.

To facilitate the arguments which follow, we shall give a linear form to Equations (33) and (34), that is:

$$q_{t+1} - \hat{q} = k_1 d_t + k_2 (q_t - \hat{q}), \tag{36}$$
$$g_t = g_t^e + k_3 d_t + k_4 (q_t - \hat{q}), \tag{37}$$

where the four coefficients k_1, k_2, k_3, k_4 will be positive and k_2 smaller than 1. Here, as in many other macroeconomic analyses, the main intention is to study the effects of relatively small changes, for which linearity is allowable. Nevertheless, we should remember the underlying non-linearity: if utilization of capacity is particularly intensive, then k_1 must be so small as to be negligible, and k_4 must on the contrary be large. In the opposite case of very low capacity utilization, k_1 must be close to 1.

For more common situations in which q_t is close to the normal $\hat{q}$, we may wonder about the likely relative positions of the four coefficients[43]. For that,

[43] The reader might find it interesting to consult J. Tobin, Keynesian models of recession and depression, *American Economic Review* (May 1975), where the models are quite comparable to the one studied here. J. Tobin refers to a "Marshall model" where $k_1 = k_4 = 0$. He makes a detailed study of what he calls the "Walras, Keynes and Phillips model" where $k_2 = k_3 = 0$.

it is interesting to consider the case in which demand would correspond to a just normal utilization of capacity ($D_t = \hat{q}\bar{y}$) and only inertia of productions would lead to a different degree of utilization; then production ought to evolve towards the normal capacity utilization, which implies $k_2 - k_1 > 0$, and inflation ought to exceed, or fall short of, expectations depending upon whether production would be higher or lower than that normal level, which implies $k_4 - k_3 > 0$. On the whole, we may thus set the following inequalities, which will be used in the arguments:

$$0 < k_1 < k_2 < 1, \qquad 0 < k_3 < k_4. \tag{38}$$

2.2. The dilemma: stagnation or inflation

1. Current ideas about short-term adjustments in prices and productions, as well as on the formation of expectations about inflation, lead to believe in a dynamic dependence going from demand towards production and the price level. This dependence ought to present somewhat permanent properties and would be responsible for recurring effects shaping short-run business trends. The present section is devoted to a problem resulting from it.

If $\{d_t\}$ represents the sequence of values taken by pressure of demand and if, likewise, $\{p_t\}$ and $\{q_t\}$ represent the sequences concerning the price level and the degree of capacity utilization, the assumptions made in the previous section indeed imply the two causal links illustrated by Figure 2. It is precisely these links that we have to examine.

We begin here only with the first stage of our theoretical elaboration. Indeed, whilst the pressure of demand depends on exogenous factors, including

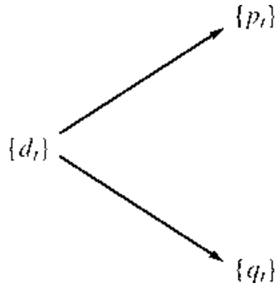

Fig. 2.

economic policy, it also depends on the way in which prices and productions evolve. There is, therefore, feedback from $\{p_t\}$ and $\{q_t\}$ to $\{d_t\}$, which we shall have to take into account later for a fuller theory. Ignoring the feedback in a first stage is interesting, since this suffices to bring out very well the dilemma facing any economic policy acting exclusively on the aggregate demand for goods and services: either to reduce demand, thereby obtaining the favourable result of slowing down inflation, but then depressing output; or to stimulate demand, thereby inducing more output, but then inflation will accelerate.

This dilemma arises because the logical structure of the links illustrated in Figure 2 results in a stable relationship between the time path of the price level $\{p_t\}$ and the time path of output $\{q_t\}$. To get away from this relationship, one ought to act otherwise than through the demand for goods and services. We shall come back to this later; but for the moment we must examine the dilemma closely.

The stable relationship can be derived even before we specify how expectations are formed. Indeed, elimination of d_t from (36) and (37) leads to:

$$k_1\left(g_t - g_t^e\right) = k_3\left(q_{t+1} - \hat{q}\right) - k_5\left(q_t - \hat{q}\right), \tag{39}$$

where

$$k_5 = k_2 k_3 - k_1 k_4. \tag{40}$$

We shall directly use this form of the constraint in the next part of the chapter. We now want to characterize further the dynamic process for the case of extrapolative price expectations.

2. The dynamic system is made up of Equations (36) and (37) together with the equation which defines the sequence $\{g_t^e\}$ in terms of the sequence $\{g_t\}$. Rather than to keep Equation (6) at this stage we are going to use the following slightly more general equation, which is appropriate when the stochastic process perceived by the agents as ruling prices is given by (9), namely:

$$g_t^e = a(1-b)\sum_{\tau=1}^{\infty} b^{\tau-1} g_{t-\tau} + (1-a)\hat{g}. \tag{41}$$

The positive number a will be exactly equal to 1 if Equation (6) applies: if not, it will be smaller, which means that agents will expect some reversion towards a long-run inflation rate $\hat{g}$ (the phrase of “regressive expectations” is then sometimes used, as was said in Section 1.4).

Moreover, from now on, it will be convenient to use the symbolic lag operator L, which is, by definition, such that:

$$x_{t-\tau} = L^{\tau} x_t \tag{42}$$

for any variable x_t. Rules to which its use must obey are assumed known[44]. Equation (41) is thus written:

$$g_t^e - \hat{g} = \frac{a(1-b)L}{1-bL}\big(g_t - \hat{g}\big). \tag{43}$$

Rewriting (36) and (37) with the operator L and introducing the above expression for g_t^e, we obtain:

$$(1-k_2L)\big(q_t - \hat{q}\big) = k_1 L d_t, \tag{44}$$

$$\frac{1-cL}{1-bL}\big(g_t - \hat{g}\big) = k_3 d_t + k_4\big(q_t - \hat{q}\big), \tag{45}$$

where:

$$c = a + (1-a)b \tag{46}$$

(c will be equal to 1 in the case where Equation (6) applies; it would be equal to b if short-term price expectations were identical to long-term ones and could so be assumed exogenous in our present discussion).

Equation (40) corresponds to the lower arrow in Figure 2; the upper arrow corresponds to the equation which we deduce from (45) when replacing $q_t - \hat{q}$ by its value, that is:

$$(1-k_2L)(1-cL)\big(g_t - \hat{g}\big) = (k_3 - k_5L)(1-bL)d_t. \tag{47}$$

Finally, the stable relationship between $\{g_t\}$ and $\{q_t\}$ is deduced from (44) and (45) when we eliminate d_t, that is:

$$k_1(1-cL)\big(g_{t-1} - \hat{g}\big) = (k_3 - k_5L)(1-bL)\big(q_t - \hat{q}\big). \tag{48}$$

[44] The main point is to know how functions of L, such as $F(L)$, should be interpreted. This is obvious if $F(L)$ is a polynomial. We shall also consider rational functions, such as the ratio in Equation (43). Then the function will stand for its power expansion, often an infinite sum. The function multiplied by a variable, such as g_t, will then represent an infinite polynomial, which will always be here assumed to converge.

We have to draw the consequences of the three Equations (44), (47) and (48). Before anything else we note that, if capacity utilization is low, k_1 must be high and k_3 low; in such a case a large decrease in output has to be accepted in order to really slow down inflation by demand management. The situation is more favourable in the case of a high utilization of capacity, for then k_1 is low and k_3 may be high.

3. Let us now focus on the case in which k_5 is zero, which is compatible with inequalities (38) and should be sufficient at this stage of our theoretical examination (we shall, nevertheless, remember that k_5 should tend to be positive if capacity utilization is high, but negative in the opposite case, such properties resulting from the non-linearity of functions φ and π). With $k_5 = 0$ Equation (48), expressing the dilemma confronting aggregate demand policy, assumes a simpler form. Let us look at the two cases concerning price expectations.

With purely adaptive expectations ($c = 1$), the dilemma is expressed by:

$$g_t - g_{t-1} = \frac{k_3}{k_1}(1-b)\big(q_{t+1} - \hat{q}\big) + \frac{k_3 b}{k_1}(q_{t+1} - q_t). \tag{49}$$

The acceleration of inflation is positively linked to, the one hand, the degree of capacity utilization, on the other hand, the change in this degree. *If capacity utilization fluctuates relatively slowly around its normal value, then the rate of inflation will also fluctuate, but with a lag*: this rate will be highest when utilization has already started to decrease in the direction of its normal value[45].

In the case of regressive expectations ($c < 1$), Equation (48) implies:

$$g_t = \hat{g} + \frac{k_3}{k_1}\sum_{\tau=0}^{\infty} e_\tau \big(q_{t-\tau+1} - \hat{q}\big), \tag{50}$$

where coefficients e_τ are positive and defined by:

$$e_0 = 1, \qquad e_\tau = (c-b)c^{\tau-1} \quad \text{for } \tau > 0. \tag{51}$$

The gap between the rate of inflation and its normal value also depends positively on the past and present gaps between the degree of utilization and its normal value; g_t is therefore lagged with respect to q_t, but the lag, which increases with c, is shorter than in the previous case. Note moreover, that with the

[45] We should remember that g_t is the relative change between p_t and p_{t+1}; it materializes at the same time as q_{t+1}.

general expression (48), the lag will also be a decreasing function of k_s/k_3 but cannot completely disappear[46]. Thus it tends to be particularly short in times when the degree of capacity utilization tends to be high.

4. This lag, which has often been established empirically, is linked to another observed fact, namely *that any reduction in aggregate demand has first a depressive effect on output before having any clearly noticeable effect on the price level.*

To illustrate this property of the dynamic system, consider the following hypothetical case: a balanced inflationary growth has been achieved until period $t = 0$ (that is $q_t = \hat{q}$, $d_t = 0$ and $g_t = g_0$ for $t \leqslant 0$); with the aim of slowing down inflation, a negative demand pressure is enforced $d_1 = -\varepsilon$ during period $t = 1$; but demand pressure returns to being nil again later ($d_t = 0$ for $t > 1$). It is not difficult to calculate what the evolutions of capacity utilization and the rate of inflation will be. Equation (44) shows that output is depressed from period $t = 2$ on, the sharpest fall occurring precisely in period 2; in the long run the rate of utilization progressively moves back towards its normal value $\hat{q}$:

$$q_1 = \hat{q}, \qquad q_t = \hat{q} - \varepsilon k_1 (k_2)^{t-2} \quad \text{for } t \geqslant 2. \tag{52}$$

Likewise, Equation (47) determines the evolution of the rate of inflation. The solution is somewhat complex in the general case. It is given here only for the case of purely adaptive expectations ($c = 1$). We easily calculate:

$$g_1 = g_0 - \varepsilon k_3, \qquad g_2 = g_1 - \varepsilon (k_1 k_4 - b k_3) \tag{53}$$

and for $t \geqslant 3$:

$$g_t = g_{t-1} + \varepsilon k_1 k_4 (b - k_2)(k_2)^{t-3} \tag{54}$$

in the long run the rate of inflation tends towards the limit:

$$g_\infty = g_0 - \frac{\varepsilon (1-b)(k_3 - k_5)}{1 - k_2}. \tag{55}$$

There is a permanent slowing down of inflation ($k_2 < 1$ implies $k_5 < k_3$). If $b > k_2$ and $k_5 = 0$, deceleration turns out to be concentrated in the first period

[46] In the limiting case, considered by J. Tobin, where $k_2 = 1$ and $k_3 = 0$ and where price expectations are purely exogenous ($a = 0$), then (48) reduces to a proportional relationship between $g_t - \hat{g}$ and $q_t - \hat{q}$; taking into account the definition of g_t, the rate of inflation is still lagged with respect to the degree of utilization.

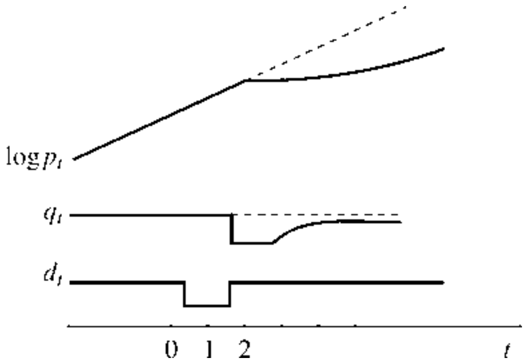

Fig. 3.

(that is, for the increase in prices between $t = 1$ and $t = 2$); later, the rate of inflation slightly rises, but without coming back to its initial value g_0.

Time paths of the three variables d_t, q_t and p_t are shown in Figure 3. We there see why the depressive measure may be said to have a more immediate impact on output than on prices.

2.3. A first look at empirical confirmation

The model of short-run adjustments, developed in the two foregoing sections, has observable implications. This is why we now stop for a moment in order to see how these implications stand when confronted with actual macroeconomic time series. But we cannot expect such a simple model to perfectly fit the data. Common sense suggests that the dynamics of actual phenomena is more complex: it certainly reacts to changes in determinants which were ignored above; it involves more lags that assumed, for instance, in Equations (36) and (37), etc. Moreover, there is a feedback from the evolution of prices and quantities to the aggregate demand for goods and services. The feedback, which will be studied in the next section, actually interferes with in the observed dynamics. In other words, the model is not yet complete and ready for a fully revealing confrontation with the facts. This section, literally speaking, provides just a first look, but one which we ought to keep in mind later on when fuller but less transparent tests and estimates will be discussed.

1. Let us consider the implied relation between changes in the speed of inflation and changes in output, as expressed by an equation such as (48). The

empirical equivalents to the abstract variables g_t and q_t ought to be detrended since we are concerned here with the short-run dynamics. The variable q_t might then be obtained from the time series giving the volume of GDP. Here we shall at first remain close to the initial presentation of Section 2.1 and rather consider series of the degree of capacity utilization. Existing series of that type do not give figures for the whole economy, only for the industrial or manufacturing sector; but they appear to provide a faithful indicator for short-run movements of output in the non-agricultural economy (the output of services exhibits only a very low variability around its trend). More delicate is the question of knowing whether the concept measured in the statistical series is the most relevant one for a test of our model.

The series comes from surveys in which firm or plant managers are asked to give figures about their capacity utilization, meant to be a ratio of the actual level of output to a sustainable maximum level. Although the concept is not precisely defined but rather explained, managers do not find it difficult to answer. Capacity is not everywhere determined from the same elements, as it depends in particular on the technology and organization of production, but for managers it makes sense to speak of how much their facilities can produce without extraordinary efforts[47]. Clearly this differs from what would correspond to the macroeconomic concept of full employment output, to which we alluded at places in Chapter 7; it rather refers to what existing equipments and infrastructures would permit at a short-term horizon.

It is indeed observed that acceleration of inflation and capacity utilization move closely together, the series of the second usually moving in advance of that of the first, exactly the implication of Equation (49). As one might expect the correlation is highest (with a coefficient of roughly 2/3) when inflation is measured from a producer price index for finished and intermediate industrial goods. But the correlation is still high when reference is made to what is commonly called "core inflation" (measured from the consumer price index excluding food, energy, perhaps also incidence of tax changes).

Fits to an equation relating quarterly changes in the rate of inflation to current and lagged degrees of capacity utilization have also been found interesting, particularly in comparison with fits to better known equations where the right-hand member contains the rate of unemployment rather than the degree of capacity utilization (in Part 6 we shall consider at length the latter equations)[48].

[47] On this and what follows, see C. Corrado and J. Mattey, Capacity utilization, *Journal of Economic Perspectives* (Winter 1997).

[48] Overall, capacity utilization is found to be a better leading indicator of changes in inflation than the rate of unemployment. See W. Franz and R. Gordon, German and American wage and price dynamics, *European Economic Review* (May 1993); D. Staiger, J. Stock and M. Watson, The

A good fit requires introduction of a few additional exogenous variables such as changes in food prices and in import prices, or changes in factors acting on wage formation such as union militancy. It also requires a less narrow dynamics than that of Equation (49) (see, for instance, Franz and Gordon, op. cit.).

2. Another type of empirical confirmation comes from the study of available annual series of the inflation rate and output over long periods. Although accuracy in estimation of the dynamic profiles would require quarterly series, which are not available before the middle of this century, a long historical perspective is interesting. Since surveys on capacity utilization were not made in earlier decades, proxy for the ratio q_t must be found, for instance, by deriving from the series of output an estimate of capacity; this is sometimes simply done by exponential interpolation between cyclical peaks of the output series; the study to be now reported rather brings to output a correction built from unemployment of the labour force excluding self-employed farmers and proprietors.

The study in question, made by R. Gordon[49], bears on the United States for the period 1892 to 1978. Examination of the series suggests making a distinction between several periods. Thus, from 1892 to 1929, apart from 1915 to 1922, the estimation could be expressed approximately, in our notations and neglecting a constant term by:

$$g_t = 0.1 g_{t-1} + 0.4 q_{t+1} - 0.1 q_t - 0.1 q_{t-1}. \tag{56}$$

The equation is close to that derived from (48) with $b = c = 0.1$, $k_3/k_1 = 0.4$ and $k_5/k_1 = 0.05$.

Likewise, from 1953 to 1978, and apart from effects of the price freeze imposed by the Nixon administration, and of exceptional changes in the prices of energy and food, the estimation leads approximately to:

$$g_t - g_{t-1} = 0.2\big(q_{t+1} - \hat{q}\big) + 0.3(q_{t+1} - q_t), \tag{57}$$

which would correspond to (49), thus also to (48), with $b = 0.6$, $c = 1$, $k_3/k_1 = 0.5$ and $k_5/k_1 = 0$. (We must remember that in Equations (56) and (57) the unit time period is a year, not a quarter.)

NAIRU, unemployment and monetary policy, *Journal of Economic Perspectives* (Winter 1997).

[49] R. Gordon, A consistent characterization of a near-century of price behavior, *American Economic Review* (May 1980).

2.4. The formation of demand

1. For the study of inflation it is not enough to know how changes in pressure of demand affect output and prices. These changes are at times exogenous to the phenomenon under examination; we then speak of "demand shocks". But, whether for the study of the final impact of such shocks or for the answer to questions concerning the effects of changes in the instruments of economic policy, we have to know how behaviours also affect the pressure of demand. This is fairly easy after the material presented in Chapter 7; but we now need to formalize its implications for the short-run dynamics. Such is the object of this section. Let us begin, however, with a quick comment about the introduction of policy instruments in the models.

Formalizing the initial impact of fiscal policy on the pressure of demand is straightforward: the policy is sometimes to directly reduce government demand, sometimes to squeeze disposable income with a view to reducing the demand of private agents. Certainly, the policy can be somewhat thwarted by induced effects which lead to increase in other components of aggregate demand. But we know how to study such effects. We know in particular that they depend on the type of control exercised by monetary authorities on financial operations, hence also on what are the instruments of monetary policy.

However, precisely going from a static to a dynamic analysis in which prices are endogenous is less straightforward when we wonder how the pressure of demand reacts to interventions affecting the nominal interest rate r_t or the money supply M_t. We have indeed to take into account simultaneously the feedback exerted by the evolution of prices and productions on the pressure of demand. We thus have to examine relationships of the type shown in one of the two diagrams of Figure 4. In this section, which is still preliminary, the adjustment process illustrated in Figure 2 will play no role. Part 4 of the chapter will set up and use the full model, so combining the elements exhibited in Figures 2 and 4.

2. The behaviours of households and firms play the main role in forming aggregate demand. We studied them enough to now deal with them quickly. We have to remove two simplifications accepted in Chapter 7: since the analysis was static, it neglected lags with which behaviours operate; in addition expectations were taken as exogenous. Now we have to introduce lags and to specify how, in each period, expectations depend in particular on previous evolution.

In this section and in the two first ones of Part 4, we shall hold the hypothesis that expectations at each date can be expressed as unchanging functions of past and current values of economic variables. We shall even limit attention to price expectations. As we saw in Section 1.5, the hypothesis covers cases in which

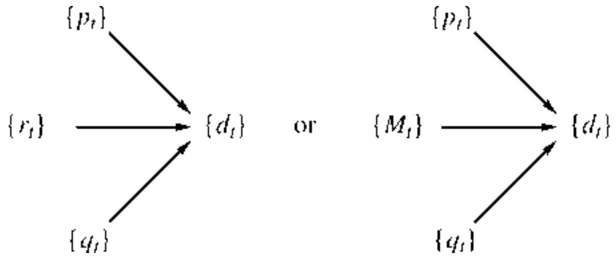

Fig. 4.

agents rationally form their expectations on the basis of a simple model of the determination of prices. But the hypothesis will not cover sophisticated retroactions of policy on expectations; for instance, a deliberate change in the interest rate, the money supply or public expenditures will not be interpreted by private agents as announcing what future policies might be. For clarity we shall speak of adaptive expectations, even though the case will be more general than a naive vision of the concept might suggest. In Part 3, in Section 4.3 and at places later, we shall consider also forms of the so-called rational expectations.

The equilibrium of the goods market led in Chapter 7 to an expression of aggregate demand in terms of the volume of production, the price level and the interest rate (see equations at the beginning of Section 1.5 of Chapter 7). In the same way, here, d_t will depend on the sequences of the values $\{q_t\}$, $\{g_t\}$ and $\{r_t\}$. This dependence is obviously complex. Any theory in this respect is confronted with two opposing risks, that of simplifying too much and hence ignoring important aspects of phenomena, and that of remaining at such a level of generality that even fairly robust aspects cannot be captured. Here we shall not aim at elegant models because this would mean eliminating too much of the complexity of the phenomena; we must be aware of the many features involved in a full understanding of the facts, and thus not baulk at examining rather daunting models. Nevertheless, we shall deliberately accept two sorts of simplification.

On the one hand, we shall keep strictly to linear expressions even though we learned, in Part 4 of Chapter 7, of significant departures from linearity which may be involved in the formation of demand. So we shall end up with an equation of the following sort for the pressure of demand:

$$d_t = A_q(L)q_t + A_r(L)r_t + A_p(L)g_t + A_t, \tag{58}$$

where $A_q(L)$, $A_r(L)$ and $A_p(L)$ are rational functions of the operator L (in Equation (43) and those following we have already introduced various rational functions of this sort). In the same way we shall choose a linear equation when we shall have to bring in the demand for money. Linear specifications are convenient because they can be easily combined with a linear model of the adjustment process and because they ease the analysis of dynamic properties of the resulting process. Although disputable, linear specifications will suffice for our study of the short-run dynamics (we have seen in Chapter 6 how involved the dynamics may be in non-linear models; we shall have to consider some such models in Chapter 9).

On the other hand, we shall stick to the simplest models available to show the main lags affecting household consumption, firms' investment, the demand for money and the adaptation of expectations. These specifications will be supposed to give a rough approximation, but not to faithfully reproduce the sequences of coefficients included in econometric equations with distributed lags. They will suffice for a theoretical exploration. In the next chapter we shall examine how applied economic studies can take account of more adequate models.

3. Before considering again the consumption and investment functions, we must realize that the study of inflation may lead us to see in a new light the relation between the pressure of demand and prices.

In the analytic framework of Chapter 7, the increase in the price level had a depressive effect. The size of this effect was shown by the ratio a_p/a_y of the coefficients defined by Equations (28). The depressive effect, that is the sign of a_p, was mainly due to two terms in the expression: X, the value of government demand, and Sf'_Y, approximately equal to the value of household demand induced by the net transfer payments S they received. A price increase was supposed to produce a reduction in the volume of government demand and a fall in the purchasing power of transfer payments received by households, thus a reduction in the volume of their consumption.

But if agents are sensitive to inflation, they will see to it that public budgets are rapidly adjusted in terms of price increases actually observed. The volume of public consumption and the purchasing power of transfer payments will be maintained mainly thanks to rapid revisions in the budgets initially adopted. If awareness of inflation is really acute, flows in value terms will be practically indexed to anticipated price increases, which will moreover adapt very quickly to the observed rate of inflation.

For modelling such a situation, we should no longer take X, S and τ for the exogenous variables representing the fiscal instruments, but rather G, S/p

and τ. If we make substitution in Equation (27) of Chapter 7, two positive terms, X and Sf'_Y, disappear from the expression for a_p.

Moreover, when price expectations quickly react to the current rate of inflation, a given nominal interest rate corresponds to a real rate, which is all the lower as prices rise faster. Higher inflation then induces higher demand, at least on this account: in their investment and inventory decisions firms will realize that the real cost of capital is particularly low; there is even a risk that consumers will overspend their money holdings.

So we see before us a whole range of situations between two extremes:

- the case of an economy which is not concerned by inflation, an economy where price expectations are stable, an economy in which public budgets are fixed in nominal values and money is a safe asset. This is the case considered in Chapter 7; we saw that a price increase could then contain excess demand;
- the case of an economy which is very sensitive to inflation, in which budgets are practically indexed and any acceleration of inflation also accelerates demand for investment and inventories by all agents. In such a case, it is very doubtful whether excess demand could be automatically contained by inflation; a high increase in the nominal rate of interest, or some other counteraction, seems to be then required.

We must certainly keep in mind this distinction when we shall now choose models in order to represent the behaviour of the two main economic agents.

4. In order to keep our expressions simple, we shall not introduce any disaggregation either in the distribution of incomes, to which we shall come back in Part 7, or in the distribution of financial assets.

For households, we shall thus write the "consumption function" as:

$$C_t = a_1 \frac{1-\gamma}{1-\gamma L} y_t + a_2 \bar{y}\left(g^e_{t-1} - g_{t-1}\right) + a_3 \bar{y}. \tag{59}$$

The first term is reminiscent of Equations (254) an (255) of Chapter 2, where they provided an expression of the permanent income hypothesis; the term here expresses effects of the lags in the "production-income-consumption" chain. The second term shows the influence of discrepancies between what is expected and what is actually realized for prices, an influence which is mainly the result of lags in the adaptation of some transfer payments and incomes to changes in the rate of inflation. This term is meant to also capture the other inflationary effects linked to an increase in the expected rate of price increase, as well as other depressive effects linked to an increase in the actual rate. Coefficients a_1 and a_2 should be positive, a_1 being smaller than 1.

For investment we shall keep a simple expression of the accelerator theory but assume that the real interest rate belongs to the determinants of capital intensity. If K_t^* is the capital desired at date t, we write:

$$K_t^* = \bar{y} \cdot \left(a_4 \rho_t^e + a_5 q_{t-1} + a_6\right), \tag{60}$$

where ρ_t^e is the expected real interest rate, equal to:

$$\rho_t^e = r_t - g_{t-1}^e. \tag{61}$$

The coefficient a_4 is meant to be negative and a_5 positive. Taking lags into account we can set:

$$\begin{aligned} I_t &= (1-\nu)\left(K_t^* - K_t\right) + \delta K_t, \\ I_t &= K_{t+1} - K_t + \delta K_t, \end{aligned} \tag{62}$$

where ν is a positive number lower than 1, whereas δ is the rate of depreciation. We can write the second equation as:

$$LI_t = \left[1 - (1-\delta)L\right] K_t.$$

So, eliminating K_t in system (62), we obtain:

$$I_t = \frac{(1-\nu)(1-\alpha L)}{1-\nu L} K_t^* \tag{63}$$

an equation where α is by definition $1-\delta$. Substituting (60) in place of K_t^*, we see how investment is assumed to depend on the time path of ρ_t^e and of q_t, thus finally on the time paths of the interest rate r_t, of the inflation rate g_t and the degree of capacity utilization q_t.

Considering now the sum $C_t + I_t - y_t$, grouping under the term A_t the additive constants and the importance $G_t/\bar{y}$ of public consumption[50], finally taking into account expression (43) for g_t^e, we can identify as follows the functions $A_q(L)$, $A_r(L)$ and $A_p(L)$ of (58):

$$A_q(L) = a_1 \frac{1-\gamma}{1-\gamma L} + a_5 \frac{1-\nu}{1-\nu L}(1-\alpha L)L - 1, \tag{64}$$

$$A_r(L) = a_4 \frac{(1-\nu)(1-\alpha L)}{1-\nu L}, \tag{65}$$

[50] The exact expression of A_t is $G_t/\bar{y} + a_3 + a_6\delta - [A_p(1) + A_r(1)]\hat{g}$.

$$A_p(L) = \frac{-[a_2(1-\nu L)(1-cL) + aa_4(1-\nu)(1-b)(1-\alpha L)L]L}{(1-\nu L)(1-bL)}. \tag{66}$$

5. Let us briefly examine the sort of dynamic dependence which these expressions imply. To do this, we just need to consider the power expansions of the three functions of L defined by $A_q(L)$, $A_r(L)$, $A_p(L)$. We know indeed that, under convergence conditions here satisfied, a rational function $A(L)$ can be given a power expansion:

$$A(L) = \sum_{\tau=0}^{\infty} \alpha_\tau L^\tau,$$

where the α_τ are suitably chosen numerical coefficients. These coefficients characterize the time dependence between x_t and $A(L)x_t$, since:

$$y_t = A(L)x_t = \sum_{\tau=0}^{\infty} \alpha_\tau x_{t-\tau}. \tag{67}$$

If, for example, x_t is zero for all t except for $t = t_0$ and equal to 1 at that time, the value taken on by y_t is zero for $t < t_0$ and equals α_τ for $t = t_0 + \tau$ with $\tau \geqslant 0$.

As for the three functions (64) to (66), elementary calculations lead to the values of the series for the respective coefficients, that is:

$$\begin{aligned} \alpha_{0q} &= a_1(1-\gamma) - 1, \qquad \alpha_{1q} = a_1(1-\gamma)\gamma + a_5(1-\nu), \\ \alpha_{\tau q} &= a_1(1-\gamma)\gamma^\tau - a_5(1-\nu)(1-\nu-\delta)\nu^{\tau-2}, \quad \tau \geqslant 2, \end{aligned} \tag{68}$$

$$\begin{aligned} \alpha_{or} &= a_4(1-\nu), \\ \alpha_{\tau r} &= -a_4(1-\nu)(1-\nu-\delta)\nu^{\tau-1}, \quad \tau \geqslant 1, \end{aligned} \tag{69}$$

$$\begin{aligned} \alpha_{op} &= 0, \qquad \alpha_{1p} = -a_2, \\ \alpha_{2p} &= a(1-b)\big[q_2 - q_4(1-\nu)\big], \\ \alpha_{\tau p} &= a(1-b)\big\{a_2 b + a_4(1-b)\big[(\tau-2)(1-\delta-b) - b\big]\big\}b^{\tau-3} \end{aligned} \tag{70}$$

(the last equality holds when $\tau \geqslant 3$ and when $\nu = b$, the expression being even more complex in general).

So, the series for the coefficients $\alpha_{\tau q}$ characterizes how the pressure of demand reacts after a temporary increase in the rate of capacity utilization. The first term is negative (the marginal propensity to consume a_1 is lower than 1); the second term is positive, as are doubtless some of the following terms; but, since it is natural to assume $\gamma < \nu$, the terms $\alpha_{\tau q}$ become negative, albeit small in absolute value, when τ is large. This sort of alternating profile is due to the combined interplay of the multiplier and accelerator. We shall examine this more closely in the following chapter.

The series for the coefficients $\alpha_{\tau r}$ shows that the immediate effect of a temporary increase in the interest rate is to reduce the pressure of demand ($\alpha_{or} < 0$); but the lagged effect is the opposite ($\alpha_{\tau r} > 0$ for $\tau > 0$).

A temporary increase in the rate of inflation has, first of all, a depressive effect ($\alpha_{1p} < 0$). But the impact on the expected rate of inflation later on produces a stimulation, which lasts for one or more periods, before a depressive, if weak, effect reappears ($\alpha_{2p} > 0$ and $\alpha_{\tau p} < 0$ for high τ).

We can also follow the effect of a permanent increase in each of the three variables. In general, Equation (67) shows that, if x_t is equal to zero for $t < t_0$ and to 1 for all $t \geqslant t_0$, then the value of y_t for $t \geqslant t_0$ is equal to the sum of the $t - t_0 + 1$ first terms in the series of coefficients α_τ; this value then converges with time towards the sum of all α_τ, which may be calculated directly as $A(1)$, that is, after a pure deletion of the operator L in the equation defining $A(L)$. So, we calculate here;

$$A_q(1) = a_1 + a_5\delta - 1, \tag{71}$$

$$A_r(1) = a_4\delta, \tag{72}$$

$$A_p(1) = -(1-a)a_2 - aa_4\delta. \tag{73}$$

We might say that these formulas give the "long-run effects" of permanent increases in the variables they concern. But we should not be misled by the expression: the theory just intends to give a good account of short-run consequences; it may badly describe effects induced in the long run. In order to correctly study the latter effects we ought to refer to growth theories. Note, for example, that productive capacity $\bar{y}$ is here considered as given and constant, whereas in reality it is obviously influenced by investment. Considering here calculated "long-run effects" of permanent changes is rather a way of appreciating the types of trend that the theory implies for developments covering a certain number of periods (12 or 20 quarters, for example).

Given this proviso we may look at the long-run effects implied by Equations (71) to (73). A permanent increase in production ought to eventually lead to a fall in the pressure of demand (it is natural to assume $a_1 + a_5\delta < 1$); the result of a permanent increase in the nominal interest rate ought to be similar. Over time a permanent increase in the rate of inflation ought to have the opposite effect on the real interest rate, hence to increase the pressure of demand, except in the case of regressive expectations; apart from this case, we must indeed set $a = c = 1$ in Equation (73).

6. In a complete study of the effects of economic policies we should obviously consider the adaptation of the demand for money to the supply of money. Hence we should specify the money demand function.

Here, we shall stick to two basic notions. In the first place the purchasing power of cash holdings at the end of period t, namely M_{t+1}/p_t, ought to vary approximately in proportion to "permanent income", which has already featured as the multiplier of a_1 in the consumption function (59). In the second place, this purchasing power ought to be a decreasing function of the opportunity cost of holding money, that is, both of the nominal interest rate r_t, which concerns the trade-off between holding money and holding bonds, and of the expected rate of inflation g_t^e, which matters in the choice between saving money and acquiring goods.

So, we shall introduce the following variable, close to the usual definition of the "liquidity ratio in the economy":

$$m_{t+1} = \frac{M_{t+1}}{p_t \bar{y}}. \tag{74}$$

For the sake of simplicity we specify again here a purely linear function:

$$m_{t+1} = \frac{b_1(1-\gamma)}{1-\gamma L} q_t + b_2 r_t + b_3 g_t^e + b_4, \tag{75}$$

where b_1 is positive, whereas b_2 and b_3 are obviously negative.

Taking account of expression (43) for g_t^e we can rewrite this equation in a form analogous to that of Equation (58), that is:

$$m_{t+1} = B_q(L) q_t + b_2 r_t + B_p(L) g_t + B_0 \tag{76}$$

with:

$$B_q(L) = \frac{b_1(1-\gamma)}{1-\gamma L}, \tag{77}$$

$$B_p(L) = \frac{a b_3 (1-b) L}{1-bL}, \tag{78}$$

$$B_0 = b_4 + b_3(1-a). \tag{79}$$

In a stationary state, that is when the variables q_t, r_t, and g_t have constant values, the liquidity ratio m_{t+1} is also constant; that is, the money supply increases at the rate of inflation. If there is an unexpected acceleration in the rate of inflation with neither q_t nor r_t being affected, two opposite effects in the value M_{t+1} of desired cash holdings appear; a growing need for money to cope with more expensive operations, but also an increasing opportunity cost which discourages holding cash. It is natural to assume that the first effect dominates in the short run; we should thus impose:

$$-ab_3(1-b) < m_{t+1}. \tag{80}$$

The discussion of formulas given throughout this section could be extended. More flexible forms could be introduced so as to make the specifications more general, and this will indeed have to be done when we shall turn to closer econometric tests of the theory. But we already have a basis from which dynamic analysis of the macroeconomy can start.

3 About a Fictitious Direct Control of Inflation and the Credibility of Policies

From the various elements brought out in the foregoing part we can build a dynamic study of demand management policies in the same spirit as the static study was built in Chapter 7. Indeed, we shall do it in the next part. But we now stop for a pause so as to introduce, in a very simple framework, some important themes that economic policy discussions have uncovered during the past twenty five years or so. These themes have to do with the interaction between the choice of economic policies by government and the beliefs or mood, hence the behaviour, of private agents. In order to deal with them, the formalization has to recognize a new dimension of interdependence, hence to become more complex, unless it makes room for that dimension by fiercely simplifying otherwise. Simplification is all the more in order since the discussion dragged macroeconomists into the field of political economy, as we shall see.

This is why in this part we shall make the assumption that government directly and exactly controls the rate of inflation. It is so unrealistic and bold an assumption that few economists can easily adopt it[51]. They have to be convinced that, first, it is an enlightening pedagogical device, and that, second, results will never be accepted at face value, but rather be seriously checked within more realistic, hence more complex, models. Indeed, in the following parts of this chapter, we shall remove the bold assumption, although still ac-

[51] "The hypothesis of perfect control of the rate of inflation on a period-by-period basis is downright silly", A. Blinder in "What central bankers could learn from academics – and vice versa", *Journal of Economic Perspectives* (Spring 1997).

cepting important simplifications. Much later in this book, actually in Subsection 4.2.5 of Chapter 9, we shall exhibit a model with which R. Clarida, J. Gali and M. Gertler were recently able to treat the various themes to be considered here. They assume a direct control of the nominal interest rate, like will be most often done in this chapter. Their results lead to roughly the same implications as those to be now drawn thanks to the pedagogical device.

The assumption of a direct control of inflation simplifies because it hides another dimension of interdependence with which a full study has to deal, namely that involving aggregate demand, as it is actually dependent on macroeconomic policy, on shocks affecting demands of private agents, on inflation and on the degree of capacity utilization, as it in turn affects these two last variables. Once he or she would have accepted the bold assumption, the theorist who would otherwise know how expectations are formed might be inclined to view an equation such as (39) as directly giving how policy determines the outcome delivered through the economic system; thus a very simple set up would follow (we shall, however, see later that so using (39) might not well fit with other hypotheses made by the theorist).

The bold assumption appeared in a few noticeable and noticed articles[52] published in the 1970s, but the real landmark, showing its full potential, was posed by Robert Barro and David Gordon[53] in 1983. Their article will be surveyed in the first few sections of this part. An important feature of the analysis will be a precise scrutiny of the rational expectations hypothesis applied to the whole private sector in the presence of supply shocks.

The subsequent literature, made of a large number of significant articles, discussed issues at the frontier between macroeconomics and political economy, a discipline which takes a positive viewpoint for considering the formation of government's actions. Concepts, like policy rules, the credibility of a policy, the commitments, reputation or discretion of a government, were clarified and discussed. This literature was surveyed by Torsten Persson and Guido Tabellini[54]. Their work will be much used here in Sections 3.5 to 3.7 for what they write about monetary policy. A brief last section of this part will attempt to assess the outcome, for macroeconomics, of the literature so surveyed; the

[52] The assumption was initially made in two steps, namely that government directly controls the quantity of money M, which directly determines the price level p. Section 1.3 explained why assuming such a direct immediate dependence of p on M is irrealistic also. Even introducing in the dependence random perturbances, meant as coming from shocks in the demand for money or the demand for goods, does not make it much more realistic.

[53] R. Barro and D. Gordon, A positive theory of monetary policy in a natural rate model, *Journal of Political Economy* (August 1983).

[54] T. Persson and G. Tabellini, *Macroeconomic Policy, Credibility and Politics* (Harwood Academic Publishers, London, 1990).

main issue will then turn out to be the respective roles of political scientists and economists in guiding economic policy choices, a subject that will be taken up more fully in Section 4.6 of Chapter 9.

3.1. A simple framework

1. The model is written with just two endogenous variables for each period, in our notation[55] g_t^e, q_t. The rate of inflation g_t is assumed to be the instrument of economic policy; according to the notation given by Equation (1), it materializes in period $t+1$. Simultaneously the economic system determines what the level of economic activity q_{t+1} has to be. Barro and Gordon specify the outcome as follows:

$$q_{t+1} = q_{t+1}^n + \alpha\left(g_t - g_t^e\right), \tag{81}$$

where α is a positive coefficient[56] and q_{t+1}^n is an exogenous variable, which we shall designate as the "normal" level of economic activity for period $t+1$: it is precisely the level that would materialize if government had chosen a rate of inflation happening to be equal to the rate expected by the private sector. Finally, this normal level is assumed to follow an autoregressive random process:

$$q_{t+1}^n = \lambda q_t^n + (1-\lambda)\hat{q} + \varepsilon_{t+1}, \tag{82}$$

where λ is positive and not larger than 1, whereas the process of the random shocks ε_t is assumed to be a white noise (the successive ε_t are independently and identically distributed with zero mean).

These assumptions may be discussed in relation with the analysis and results of Part 2 above. Equation (81) is similar to Equation (39), particularly when $k_5 = 0$, a case considered as relevant in Section 2.2. Once α has been set equal to k_1/k_3, the only apparent difference is that the constant $\hat{q}$ is replaced by the random variable q_{t+1}^n. But there is also a hidden difference, namely that (39) was read as an implication of the assumptions made about the process of simultaneous adjustments, which were said to be induced in prices and

[55] Instead of the degree of utilization q_t of the long-run capacity, Barro and Gordon speak of the unemployment rate u_t, but they state that this rate is meant to be "a convenient proxy for the overall state of real activity"; it is exactly what q_t is meant to be in this part of the chapter.

[56] The notation α is chosen here for easy reference to the Barro–Gordon article, which the reader of this part may want to consult. No confusion should occur with the notation used in Section 2.4 and in the subsequent Part 4, where α refers to the rate of depreciation δ through $\alpha = 1 - \delta$.

productions by changes in the pressure of demand. This implication, hence to some extent also the assumptions, were confirmed in Section 2.3 as providing an admissible approximation to observed short-run co-movements of the price level and output[57]; but the reading of the equation was clearly different from the one suggested here by the idea of a direct control of the rate of inflation by the government; indeed, the natural interpretation of adjustment rules (36) and (37) is to say that firms choose q_{t+1} and p_{t+1} as functions of the previous pressure of demand, the previous degree of capacity utilization and the previously anticipated rate of inflation. We shall discuss in Section 4.4 of this chapter another type of rationalization of Equation (81) and in Subsection 4.2.5 of Chapter 9, the model used by Clarida et al. which will come closer to the natural interpretation of (36) and (37) but will be more systematically forward looking.

In order to give a precise meaning to the concept of normal level of economic activity q_t^n and to interpret the stochastic process (82) we have to introduce the idea of an evolving productive capacity $\bar{y}_t$ subject to supply shocks such as an increase in labour productivity or a favourable change in the terms of trade (or the reverse for negative shocks). By q_t we designate in this part of the chapter the ratio between output and the level that productive capacity would have in period t if it would be precisely equal to its long-run trend. For simplicity let this trend be the constant $\bar{y}^*$; then $q_t = y_t/\bar{y}^*$. The actual degree of capacity utilization is $y_t/\bar{y}_t$, or equivalently $q_t\bar{y}^*/\bar{y}_t$. The normal degree of capacity utilization would instead be $\hat{q}$ (which we assume to be constant); achieving this degree would require a "normal" output y_t^n such that $y_t^n/\bar{y}_t = \hat{q}$ and a level of activity q_t^n therefore equal to $\bar{y}_t\hat{q}/\bar{y}^*$. Except for multiplication by the constant $\hat{q}/\bar{y}^*$, the normal level q_t^n so evolves like productive capacity. The stochastic process (82) therefore specifies how supply shocks are assumed to change productive capacity; the higher λ the more persistent is the effect of each shock. The close link between the normal level of activity and productive capacity moreover shows how hypothesis (81) must be understood: optimal use of capacity would be reached precisely when inflation would be as expected.

[57] Barro and Gordon, in their equation to which (81) corresponds except that u has been here replaced by q, speak of u^n as the "natural rate" of unemployment and make reference to the "Phillips curve" (their inverted comas). This suggests that (81) formalizes empirical regularities of the type surveyed in Section 2.3. But an incidental footnote refers to a theoretical model due to Lucas and leading to a Phillips curve, a model in which, however, firms are price-makers (the model will be examined in the last part of this chapter).

2. Following Barro and Gordon we shall explicitly consider the implications of the target-and-instrument approach, which was exposed by J. Tinbergen[58] for the static case and later extended to the dynamic case by H. Theil[59]. The approach was already used in this book several times, more or less loosely (but precisely in parts of Chapter 5 with optimal growth and some of its applications). It will be used again here later in recurrent references to the methodology of economic policy.

The targets of policy are, in the present case, those already identified by the dilemma of Section 2.2: to stimulate economic activity as much as possible and to avoid inflation as much as possible. Since the two objectives are conflicting with one another, there will be no solution unless we specify their respective importances. This is done by a numerical cost function whose minimization will be the aim of government. Actually, the dilemma may go farther because the government may face an intertemporal problem: the rate of inflation chosen for period t may affect the terms of the dilemma in subsequent periods, as we shall see. In such a case, there will also be a conflict between present and future objectives. For resolution of the global problem the cost function to be minimized will have to be defined over all these objectives. Finally, there are random elements in the system, the function must be defined over random variables.

Resolution might be quite difficult, and theory would be bound to remain so general as to be of little use, if the theorist did not accept to consider special forms of the cost function. Two strong restrictions are fairly common in the discussion of such dynamic stochastic minimization problems. The first specifies linear separability with respect to random events and time; even more precisely the overall cost function is written as:

$$L = E\left[\sum_{t=1}^{\infty} \gamma^t Z_t\right], \tag{83}$$

where E is the mathematical expectation operator (conditional on the information available at the initial time), γ is a positive discount factor smaller than 1 and Z_t is the random cost in period t. (This type of separability was used

[58] J. Tinbergen, *On the Theory of Economic Policy* (North-Holland, Amsterdam, 1952); *Economic Policy: Principles and Design* (North-Holland, Amsterdam, 1956).

[59] H. Theil, *Optimal Decision Rules for Government and Industry* (North-Holland, Amsterdam, 1964).

several times already in this book.) The second restriction specifies that Z_t is a quadratic function of its arguments, more precisely here:

$$Z_t = \left(kq_{t+1}^n - q_{t+1}\right)^2 + b(g_t)^2, \tag{84}$$

where the parameter b is positive and k is not smaller than 1 (the notation is chosen so as to be close to that of Barro and Gordon; b and k should not be confused with similar labels for other parameters appearing in previous and subsequent parts of the chapter).

The quadratic form given to Z_t is chosen for simplicity (together with linearity of the constraints, it implies the convenient certainty-equivalence property). It may be taken here as an admissible approximation. The parameter b is meant to capture the cost of inflation relative to that of a too low use of productive capacities (including a too high unemployment of labour). We note that a high level of activity has also a cost, according to (84); with $k = 1$ this cost would even appear as soon as the degree of capacity utilization would be larger than "optimal". In order to rationalize the assumption of an increasing cost for high values of q in such a simple little model, it would be possible to say that a higher degree of utilization than $\hat{q}$ implies in the firms excessive congestion or wear and tear costs, which are not deducted in the measure of output (if we may refer here to (37), we see that firms faced with a just nil pressure of demand begin to charge higher prices than expected precisely when q exceeds $\hat{q}$). But, taking account of the employment of labour and of the inefficiencies associated with unemployment, as well as with all implications resulting from the excess-burden of public finance, Barro and Gordon argue that overall costs start increasing only beyond a higher value of q than $\hat{q}$; they therefore pose $k > 1$ (actually $k < 1$ in their notation), an assumption which will be accepted here.

3. With the two constraints (81) and (82), and the objective function defined by (83) and (84), we still need to make a precise assumption as to the formation of price expectations. Barro and Gordon assume rational expectations, as defined by Equation (12) and discussed in Section 1.5, with *full knowledge by private agents of all what is also known by the government before its decision.* More precisely the information set J_t of all agents in period t contains the normal level of activity q_t^n for that period, as well as the actual levels of activity q_t and prices p_t and all past values of these variables. This set also contains knowledge of the constraints that will apply in the future, starting with those involving q_{t+1}^n, q_{t+1} and g_t (or equivalently p_{t+1}). It finally contains knowledge of the objective function of the government defined by (83) and (84).

For choosing the value of the price expectation variable g_t^e, which is identical with the expectation of the outcome g_t of the simultaneous government decision, private agents have to figure out what this decision is likely to be, and they are assumed to do it "rationally", that is, in full knowledge of the information set. Conversely, in order to solve its decision problem in period t, the government has to figure out what its constraints really are, hence how private agents form their expectations g_t^e. The two choices, of the expectation on the one hand and of the decision on the other, are so "interactive".

Dealing with such interactive problems is precisely the province of the theory of games. On reflection it appears fairly natural that, much more generally, the interplay between private and public decisions could often be viewed as a game. Indeed, it is enlightening to consider some problems posed by macroeconomic policy by taking the approach of game theory. Hence, discussion of the simple framework posed here should serve as an interesting introduction to all such problems.

3.2. An optimal fixed rule

The opposition between fixed and discretionary rules was introduced long ago in discussions on monetary theory. This occurred first when Milton Friedman argued for a fixed rate of growth of the quantity of money against the Keynesian counter-cyclical recommendation requesting the monetary authorities to expand (resp. contract) the money supply when their analysis of current business trends led them to forecast a low (resp. buoyant) activity[60]. Later, "discretionary" policies were also meant to refer to government actions intended to engineer unexpected rises in activity. The use of the distinction by Barro and Gordon is more subtle, as we shall see.

For the moment let us consider that government looks for the optimal rule among those which could be announced and applied with such determination that private agents would be sure of their enforcement. The rule then says what rate of inflation should be chosen given J_t, let us say $h(J_t)$. Since private agents know the rule, know that it will be applied and know J_t, they expect for sure the rate of inflation to be $h(J_t)$. So we write:

$$h(J_t) = g_t = g_t^e. \tag{85}$$

The second equality is then a constraint that government must take into account in finding out the optimal function h.

[60] See, for instance, M. Friedman, *A Program for Monetary Stability* (Fordham University Press, New York, 1960).

In our present framework the optimization problem is so simple as to look trivial. Indeed, the constraint imposed by the second equation in (85) and the constraint (81) imply that q_{t+1} will be equal to the exogenous q^n_{t+1}. Hence, in Z_t given by (84), the first squared bracket will not depend on the choice of the rule. Minimizing L is therefore equivalent to minimizing the mathematical expectation of the non-negative discounted sum of the $(g_t)^2$. This can be easily achieved, by taking $g_t = 0$ each period, a very simple fixed rule. (We note that, if in (84) instead of $(g_t)^2$ we would have introduced $(g_t - \hat{g})^2$ with an exogenous small target $\hat{g}$ for the inflation rate, the optimal fixed rule would have simply been $g_t = \hat{g}$.)

3.3. A Nash solution

Let us now take the perspective of game theory. It then appears that the optimal fixed rule is not a "Nash equilibrium" of the non-cooperative game corresponding to the interactivity between the expectation of private agents, who choose the g^e_t, and the decision of government, which chooses the g_t.

In order to fully define the game, we just have to specify the payoff function of the representative private agent playing against the government; clearly it has to be a function of the errors in expectation; let us say it is simply the sum of the discounted $-(g_t - g^e_t)^2$. By definition a Nash equilibrium is a couple of strategies of the two players, each strategy being the best response to the strategy of the other player. The sequence of expectations $g^e_t = 0$, is certainly the best response of the private agent to the sequence of decisions $g_t = 0$; but the sequence of decisions $g_t = 0$ is not the best response of the government to the sequence of expectations $g^e_t = 0$.

Its best response to such a sequence would then not disregard the opportunity of stimulating somewhat activity at the cost of some inflation. Indeed, in its objective function (83), Z_t would be given by:

$$Z_t = \left[(k-1)q^n_{t+1} - \alpha g_t\right]^2 + b(g_t)^2 \tag{86}$$

once (81) would be taken into account. Minimization of the mathematical expectation of Z_t conditionally on J_t and on (82) would lead to:

$$\left(\alpha^2 + b\right)g_t = \alpha(k-1)\left[\lambda q^n_t + (1-\lambda)\hat{q}\right]. \tag{87}$$

Of course, if this rule would be followed by the government, the best response of the private agent would not be $g^e_t = 0$.

There is, however, a Nash equilibrium to the non-cooperative game. It can be found by both the government and the private agent since they have and will have exactly the same flow of informations: the sequence of the J_t. Each can compute its best response to any strategy of the other, as well as the best response of the other to any of its own strategies. Let us consider, for instance, how the government finds its response to any given sequence of g_t^e.

By a similar argument to that leading to (87) we are led to:

$$\left(\alpha^2 + b\right)g_t = \alpha(k-1)\left[\lambda q_t^n + (1-\lambda)\hat{q}\right] + \alpha^2 g_t^e. \tag{88}$$

But the government knows that the private agent has derived the same formula, following the same argument. Hence, the private agent can reach the maximum value (zero) of its payoff by selecting $g_t^e = g_t$. The government knows that and therefore chooses:

$$g_t = \frac{\alpha}{b}(k-1)\left[\lambda q_t^n + (1-\lambda)\hat{q}\right] = g_t^e \tag{89}$$

as long as it wants to give the best response to its opponent strategy and as it believes that the private agent also wants to give the best response.

Note that in (89) the square bracket is equal to the expected value in period t of the level of activity that would correspond to the optimal use of capacity in period $t+1$; the higher this value the larger is also the chosen rate of inflation. Moreover, the latter is all the larger as are k (the cost of unemployment and other distortions associated to a low level of activity) and α (the slope of "the Phillips curve"), and as the weight of inflation (b) in the objective function is lower. Note finally that, in this Nash equilibrium, productive capacity will always be optimally used, according to (81).

3.4. Time inconsistency

In the simple framework posed by Barro and Gordon, the foregoing arguments lead to two focal solutions to the policy problem faced by a government who would control the rate of inflation while rational expectations would prevail in the private sector. Which one of these two solutions is the more persuasive as providing a positive theory of inflation?

Two routes are open for answering the question. Either you are a game theorist and you will search deeper into the logic of interactive rationality. Or you are a political scientist and you will look around for facts and ideas suggested by present and past policies. Here we shall not go far along either route; we shall rather consider the directions in which they seem to go. We shall need no more in order to realize that other positive solutions to the policy problem than the two focal ones can well be envisaged, and that the problem has several dimensions, which should be kept in mind. This will serve as an introduction to subsequent sections of this part of the book.

1. The optimal fixed rule would provide the natural solution to the game if we could rely on the idea that cooperation between policy makers and private agents would spontaneously occur. But such is not the case. On the contrary, following the fixed rule is not a robust strategy of the government with respect to the permanent temptation of stimulating, for once, economic activity. Indeed, compare [$g_t = 0$ for all t] with [$g_t = 0$ for all $t \neq 1$; g_1 positive and small, chosen for instance according to (87), but with no prior announce]. The alternative strategy would so contain an exceptional and unexpected discretionary move. If, convinced that the move was exceptional, private agents stick for $t > 1$ to their expectations $g_t^e = 0$, then the cost Z_1 will decrease, in comparison with the fixed rule strategy with no change in any other Z_t. The alternative strategy so turns out to be superior.

We then say that the fixed-rule strategy is "time inconsistent", the term meaning that the decision maker has, at some times or at all times, an incentive to deviate from the strategy when other agents expect the strategy to be followed[61]. A time-inconsistent strategy is so not robust with respect to the choices of the decision maker. Neither is it with respect to the choices of other agents, who may realize the risk of being fooled and take it into account, so making the payoff of the strategy different from what was assumed.

The lack of robustness of a fixed rule may be particularly acute in cases in which the game does not go on beyond a terminal period T. Just assume the case would apply in the particular problem discussed here. Then, in period $T - 1$ the government would have a strong incentive to make an unexpected move, since observation of the move in period T would not matter. The difficulty is not removed by the remark that private agents ought to be aware of the incentive faced by government: this awareness should indeed react on private expectations, and in a way that would deviate from what is assumed in justifications of the fixed rule. Backward induction, from the terminal to the initial period, then shows that the incentive for government to deviate from the rule is already strong in the first period.

Faced with a solution including time-inconsistent strategies the theorist naturally looks for more robust solutions. Will the Nash solution of Section 3.3, by

[61] The concept was explicitly introduced by F. Kydland and E. Prescott in Rules rather than discretion: the inconsistency of optimal plans, *Journal of Political Economy* **85** (1977) 473–490.

definition not exposed to time inconsistency, offer a more realistic alternative? Not necessarily so.

In the present example private agents rationally realize that the government and citizens have a definite preference for the fixed rule, which delivers the same levels of activity as in the Nash equilibrium and less inflation. If the Nash equilibrium must eventually be the outcome of non-cooperation, the government has a definite incentive to favour cooperation and ought to be careful before spoiling it by fooling people expectations. A private strategy with the following feature then makes sense: anticipate the cooperative result ($g_t = 0$) unless something else had occurred. For the government, aware that such a private behaviour is sensible, foregoing attempts at unexpectedly stimulating activity also makes sense.

We understand that game theorists look deeper into the problem, because what was just said does not define the threat that the government can perceive as latent from the private sector and as ready to materialize in case of unexpected stimulation: how would private expectations be exactly determined in such a case? The answer to the question requires consideration of the whole future development of the game, since there is a linkage between current policy choices and subsequent inflationary expectations. In other words, the problem must be discussed as concerning a “dynamic game” with two sequences of moves, g_t by the government and simultaneously g_t^e by the private sector, moves of earlier periods being taken into account in the decisions of both players.

2. A government ready to follow the fixed rule has all reasons for announcing it, even for declaring that abiding by the rule is a commitment towards the voters. But, is such a declaration credible? The answer probably depends on the country and the circumstances.

As against its credibility is the understandable presence of the incentive to deviate from the rule, an incentive that may have much appeal at times, for instance, when unemployment is high or when difficult elections are in sight. In many countries inflation has been a recurring problem, which past governments were not able to really solve; strong evidence is required for the public to accept the idea that an essentially new policy line has been adopted for good and for long. Trust does not spontaneously come against past lack of performance.

The credibility problem is not made easier by the fact that it is actually not obvious to identify the policy; indeed, observed evolution depends not only on policy but also on exogenous shocks, for which the government is not responsible and which cannot be avoided. This identification difficulty gives policy makers an incentive to cheat and later claim that inflation was not due to their actions. Whether they do it or not, the public suspects them to do so.

But it is also true that, in some countries, governments have a well established reputation of following permanently an anti-inflationary policy. The case of Germany comes to mind, a country which had the experience of hyperinflation in the early 1920s and again of accelerating inflation before the 1948 monetary reform; from then on, over five decades, the level of prices moved up only slowly.

Where they are systematically applied, rules restrict the manner in which policy choices are made. This may follow from accepted practices, or more formally from decisions taken at what is sometimes called the constitutional level of government (as opposed to the operating level). For instance, a deliberate design of institutions may, first, give to the central bank the sole explicit mandate of maintaining stability of the purchasing power of money and, second, leave the bank free to independently choose the means for achieving this aim. Political science might perhaps explain that the mandate is implicitly understood as implying some awareness of monetary effects on activity and that the purpose of the institutional design is to create a distance between the government, too sensitive to short-term effects, and the locus where the main dilemma of macroeconomic policy is faced.

So a number of relevant issues involve not only the concepts of fixed rule and discretion, but also those of commitment, credibility and reputation. It was quite natural that, after the article of Barro and Gordon, a number of economists discussed the meanings and roles of these new concepts. To these further developments we are now turning our attention.

3.5. Commitment, simple rules or discretion

1. The main point of departure from the hypotheses made by Barro and Gordon came from the idea that government was better informed than private agents about the aggregate state of the economy. Government observed shocks, of which the private agents were not aware; it was therefore choosing its monetary moves knowing the occurrence of those shocks and knowing that they were ignored in the private sector.

Introducing asymmetry of information was natural because questions concerning government's incentives had cropped up during the discussion. The role of incentives was at the time, and still is now, a major concern all across economic theory; it was indeed intimately connected with the introduction of asymmetrical information. Such an introduction was also required for a sharper scrutiny of the relation between government commitments and fixed rules. But for the purpose the Barro–Gordon model was not so appropriate.

This indeed explains the main difference between that model and the first, much used, model in the book of Persson and Tabellini. This new model simplifies in other respects. Let us look at it. The two Equations (81) and (82) are replaced by just one, which writes in our notation as:

$$q_{t+1} = \hat{q} + \alpha\left(g_t - g_t^e\right) + \varepsilon_{t+1}. \tag{90}$$

Formally, it looks identical with (81) and (82) in which λ would be chosen equal to zero. But the realization of the white noise process of the ε_t is assumed to be non observable by private agents choosing g_t^e, whereas the government observes it.

Another, quite secondary, difference concerns how Persson and Tabellini rationalize Equation (90). The private sector is described by a simple model of the labour market; competitive firms choose the level of activity q_t (identified in the book with employment) taking as given not only g_{t-1} but also g_{t-1}^e (identified with wage inflation); firm-specific unions set the nominal wage, hence g_{t-1}^e, then taking the expected impact on activity into account. The authors use interchangeably the two phrases "expected inflation" and "nominal wage growth", pointing to the fact that the optimal wage is changed one for one with expected inflation, according to their particular specification. The rationalization moreover permits to interpret the ε_t as supply shocks.

We may wonder whether such a rationalization is persuasive, as providing a realistic basis for a discussion of monetary policy, and whether it suffices to improve upon the weakness of the rationalizations of (81) that were proposed, a weakness which was mentioned here in the presentation of the Barro–Gordon model. But, considering the purpose and other features of the theoretical investigations reported in this part, the rationalization of (90) is a quite secondary issue. The best way to read Equation (90) is to say that a negative shock deteriorates business profitability, which can be restored by an increase in prices. Whether such supply shocks predominate in economic fluctuations will be discussed later.

Since the policy maker is now assumed to observe ε_t before choosing g_{t-1}, the choice may depend on ε_t; so, we have to analyse the problem as concerning the choice of a "state-contingent strategy" or equivalently of a "policy rule". Since moreover Persson and Tabellini keep the objective function defined by (83) and (84), simply replacing $kq_t^n - q_t$ by $k\hat{q} - q_t$, the model is "linear-quadratic", so that the optimal policy rule has to be linear with respect to the series of the ε_t (this was proved in the book of H. Theil, op. cit.). In the static analysis, which will be sufficient to start with, we may write:

$$g_{t-1} = \bar{a} + a\varepsilon_t. \tag{91}$$

Whereas the private sector does not observe ε_t when forming its expectation g^e_{t-1}, we are going to assume, again to start with, that it knows the policy rule, that is, the values of the two parameters $\bar{a}$ and a. At the end of period t, private agents will observe g_{t-1}; they will then be able to compute ε_t, but too late to take that into account.

2. The formulation now permits to give a precise meaning to the distinction between commitment and discretion. Commitment is interpreted as meaning that the government has declared it would decide on g_{t-1} before observing g^e_{t-1} and that indeed it does so. Discretion is the opposite set-up in which government observes both g^e_{t-1} and ε_t before deciding on g_{t-1}.

Notice, however, that under commitment the policy maker does not ignore the reaction of the private sector to the policy rule. On the contrary, it takes this reaction into account in choosing the values of the parameters $\bar{a}$ and a. But it will not surprise the private sector by opportunist deviations from the rule.

Let us then see what is the optimal choice of $\bar{a}$ and a in the case of commitment. Since the mathematical expectation of ε_t is zero, the expectation of private agents will be $g^e_{t-1} = \bar{a}$. Introducing in (90) this equality together with (91) and taking mathematical expectation leads to:

$$EZ_{t-1} = \big[(k-1)\hat{q}\big]^2 + b\bar{a}^2 + \big[(\alpha a + 1)^2 + ba^2\big]\sigma^2, \tag{92}$$

where σ^2 is the variance of ε_t. Minimizing with respect to $\bar{a}$ and a implies:

$$\bar{a} = 0, \qquad a = -\alpha\big[\alpha^2 + b\big]^{-1}. \tag{93}$$

The policy rule means that government will reach on average a nil rate of inflation and that its actions will mitigate the impact of shocks on economic activity; indeed (93) implies:

$$q_t = \hat{q} + b\big[\alpha^2 + b\big]^{-1}\varepsilon_t. \tag{94}$$

Mitigation will be more complete if inflation has a smaller weight in the objective function.

But the rule (93) is subject to time inconsistency. Indeed, knowing g^e_{t-1} (with (93) it is going to be nil), the policy maker can compute:

$$Z_{t-1} = \left[\alpha\left(g_{t-1} - g^e_{t-1}\right) + \varepsilon_t - (k-1)\hat{q}\right]^2 + b(g_{t-1})^2, \tag{95}$$

which is maximized with respect to g_{t-1} by:

$$\left(\alpha^2 + b\right)g_{t-1} = \alpha(k-1)\hat{q} - \alpha\varepsilon_t + \alpha^2 g^e_{t-1}. \tag{96}$$

Since we assumed $k > 1$ and since $g^e_{t-1} = 0$ with the policy defined by (93), a positive inflationary bias to the policy defined by (91) and (93) would increase the value reached by the objective function. This time inconsistency undermines the credibility of the optimal policy under commitment.

3. In order to escape this credibility difficulty, the rule ought to be chosen in such a way that the government has no incentive for opportunist deviations, which would take private agents by surprise. This requirement imposes an additional constraint on the choice of the rule. We may speak here of a "credibility constraint", even though the expression "incentive constraint", which is common in the economics of asymmetrical information, would also do.

In our case the credibility constraint directly derives from (96), interpreted as a condition for the policy rule, once g_{t-1} is given by (91) and g^0_{t-1} by $\bar{a}$ (the best response of private agents to the rule). Since the constraint must be satisfied for all possible values of ε_t, it fully determines the policy by:

$$\bar{a} = \alpha(k-1)\hat{q}/b, \qquad a = -\alpha\left[\alpha^2 + b\right]^{-1}. \tag{97}$$

The rule clearly defines a Nash equilibrium, which transposes that given by (89) in the Barro–Gordon model.

We may here speak of the rule under discretion in order to designate the result of (97). It leads to the same mitigation of the effect of shocks on activity as does the rule under commitment. In this sense we may say that the level of activity is still set optimally. But the rule has an inflationary bias, which is the cost of the credibility constraint. This bias and its cost are all the higher as the weight of inflation in the objective function is smaller.

4. The model of this section draws our attention to the fact that, in a stochastic environment, policy rules with commitment are not in general fixed rules but "state-contingent" (here (93) shows that g_t depends on ε_t). It is an important remark with respect to the discussions about monetary policy among macroeconomists.

One may object, and one has indeed objected, that in a real environment, with all its features and sources of uncertainty, contingent optimal rules and likely to be very complex. So, it may seem hard to pretend that such rules could be implemented. In practice, policy rules may have to be simple.

Such is the motivation for comparing, in the model of this section, the fixed rule $g_{t-1} = 0$ (or $\bar{a} = 0$, $a = 0$) with the discretionary rule defined by (97). If we consider the value taken by EZ_{t-1} with respectively the discretionary and the fixed rule, we find that the loss resulting from the fixed rule is smaller than that following from the discretionary rule precisely when:

$$\left[\frac{\sigma}{\hat{q}}\right]^2 < \left(\frac{x^2}{b} + 1\right)(k-1)^2 \tag{98}$$

a formula which makes sense when we consider the various parameters involved. It suggests in particular that the fixed rule is not so good when the variance of shocks is relatively large.

3.6. About the institutions of economic policy

Examination of the Persson–Tabellini model enriches the conclusions drawn from the Barro–Gordon article. To a large extent it confirms them. It also leads to pose again now a question that was already hinted at in Section 3.4: which monetary institutions are most likely to push the discretionary equilibrium policy in the direction of the more favourable commitment policy? Among the institutional devices discussed in the literature we shall here briefly consider three interesting ones: giving independence to the central bank, formally instituting a fixed rule with escape clause, favouring or barring the practice of price and wage indexation.

1. Up to now we loosely and almost systematically spoke of the government as taking decisions in matters of monetary policy. Particularly in this book where we do not pretend to discuss the inner operation of the large sector made of governmental and financial institutions, we need not remind that the central bank was responsible for implementing monetary policy. But we now introduce a political economy argument which requires the distinction between the government and the central bank. This argument indeed emphasizes the advantage of leaving the central bank decide on the monetary policy, independently of what government may say; giving independence to the bank would improve economic performance with respect to the trade-off between inflation and employment variability.

The recent literature about this argument uses the framework discussed here. Its outcome is presented in the survey of Persson and Tabellini. The main idea is that an independent central bank will give a higher weight b_B to inflation in the loss function Z_t than the weight b appearing in the true social loss function, the one that government is supposed to consider[62]. Since the equilibrium discretionary policy of the government (its policy subject to the credibility constraint) is optimal for regulating the level of activity but leads to a too high mean rate of inflation, the equilibrium discretionary policy of the bank will lead to a better result, as long as b_B will not be too high with respect to b. Indeed, substitution of b_B for b in (97) will lead to a lower mean rate of inflation; there will also be a lower degree of activity regulation, but the loss resulting from this second effect will only be second-order small in the neighbourhood of the optimal degree of regulation.

Two remarks are in order here. Firstly, the argument requires $b_B > b$. This is why the literature often speaks of a conservative central bank, or even of a conservative central banker. This might be read as having a political tone. But we must realize that the first function of a central bank is everywhere to care about the value of the currency. The positive bias of b_B with respect to b is therefore a natural consequence of the division of roles within government.

Secondly, a central bank which would care only about inflation (b_B infinitely large) would indeed avoid it (Equations (97) would lead to $\bar{a} = 0$ and $a = 0$). But, as we saw, the fixed rule imposed by the bank might be then worse than the equilibrium discretionary policy defined earlier for the government (inequality (98) gives the condition for the fixed rule to be preferable). In other words, the "optimal" central banker should not be too "conservative".

2. If difficulties of implementation rule out in practice the optimal policy under commitment, it is a natural idea to institute an explicit arrangement according to which the policy will be fixed in "normal times" but discretionary in "abnormal times". In our formulation the central bank could for instance apply the fixed rule ($\bar{a} = a = 0$) as long as $|\varepsilon_t| \leqslant e$, where e would be a given threshold; $\varepsilon_t < -e$ would provoke the trigger action of a deliberate surprise rise in inflation; $\varepsilon_t > e$ that of a surprise decrease; in both cases the bank would, according to (90), counteract the effect of an exceptionally large supply shock on the level of activity.

Finding the equilibrium discretionary policy followed when $|\varepsilon_t| > e$ is easy, because the credibility constraint (96) still applies in those cases. With a linear

[62] Within the approach here discussed the idea was first introduced by K. Rogoff in The optimal degree of commitment to a monetary target, *Quarterly Journal of Economics* **100** (1985) 1169–1190.

discretionary policy (91) then supposed to apply, there is just one difference from what was found in Section 3.5, namely that the expectation g_{t-1}^e of private agents is no longer equal to $\bar{a}$ but to $\pi\bar{a}$, where π is the probability of the abnormal event $|\varepsilon_t| > e$. Introducing this change in the application of (96) leads to:

$$\bar{a} = \alpha(k-1)\hat{q}/\left[\alpha^2(1-\pi)+b\right], \quad a = -\alpha\left[\alpha^2+b\right]^{-1}. \tag{99}$$

When the abnormal event occurs, the policy mitigates its effect on the level of activity exactly as would do the policies discussed in 3.5, but with a smaller inflationary bias.

A full discussion of this fixed rule with an escape clause must compare it with the strict fixed rule $g_{t-1} = 0$. In normal times it implies a depressing effect on the level of activity because of private expectations; indeed $g_{t-1} - g_{t-1}^e$ is then equal to $-\pi\bar{a}$. The discussion then shows that, if dispersion of the shocks is substantial and if the trigger threshold e is well calibrated, the policy with the escape clause leads to a better performance.

It may be argued that the distinction between normal and abnormal times is informally made by central bankers and that the rule just considered captures the spirit of many decisions taken by them. Rules with escape clauses were moreover explicitly codified in international exchange rate agreements, most notably in the Bretton Woods system and in the European Monetary System.

3. As we recognized in Section 1.9, the practice of price and wage indexing removes some of the costs of inflation. Indeed, the practice is widespread in high inflation countries. But it is commonly perceived as dangerous by monetary authorities, who fear that spreading indexation will entail increasing inflation. S. Fischer and L. Summers explained the logical basis of this position and, in order to do so, used precisely the approach of Barro and Gordon surveyed above[63].

The main idea is that making inflation less costly will induce the government to accept a higher rate of inflation and that it will finally be detrimental. The conclusion may be related to the formulas of Section 3.3. Indeed, Equations (89) directly show that, in the Nash equilibrium, g_t increases when the parameter b, characterizing the cost of a given level of inflation, decreases and then the cost Z_t accepted at this equilibrium also increases. But the argument deserves a closer scrutiny.

[63] S. Fischer and L. Summers, Should governments learn to live with inflation?, *American Economic Review* (May 1989).

Fischer and Summers first note that the model assumes a perfect control of inflation by the government. Since it is not a realistic assumption, we must consider that actual inflation will differ from intended inflation; the difference between the two, a random variable, may be called uncontrollable inflation. It is natural to think that this random variable is uncorrelated with intended inflation, has a zero mean and has a variance σ_u^2; so, the expected value of Z_t now has the additional term $b\sigma_u^2$, which decreases when b decreases. It is then easy to see that there is an critical value of the parameter b characterizing the cost of inflation: if b happens to be higher, reforming practices so as to mitigate the cost of any given degree of inflation is beneficial because reduction in the average cost of uncontrollable inflation will be larger than increase in the cost induced by more tolerance of inflation; the reverse applies if b happens to be lower than the critical value.

Another, probably more important, consideration comes from the idea that other outcomes than the Nash equilibrium may arise. In particular, the monetary authority may have chosen to develop the reputation for being inflation averse and for applying a low-inflation strategy somewhat similar to the fixed rule defined in Section 3.2. Then the discussion of the following section suggests that much depends on the credibility of the strategy and on how this credibility reacts to evidence that indexations, or other forms of inflation mitigation, are spreading. Fischer and Summers discuss the question and conclude that "governments whose ability to maintain low rates of inflation is uncertain should not reduce the costs of actual inflation. Measures that reduce the cost of anticipated inflation, or undercut opposition to it, are particularly likely to be pernicious".

3.7. Government's reputation

1. It is somewhat paradoxical that we could, by a purely static analysis, go as far as we did into the study of the credibility of policies. We had found appropriate to pose the objective function L as being intertemporal; but arguments up to this point considered just the component Z_{t-1} concerning period t in Equation (83), which specified L and made it linearly separable for simplicity. The concept of time inconsistency was introduced, but did not require solution of any intertemporal problem. There are, however, two important reasons for appealing also to explicitly dynamic analysis.

Firstly, when heuristically comparing in Section 3.4 the realism of the Nash equilibrium with that of the couple made of the time-inconsistent no-inflation policy and of the private expectations of a nil rate of inflation, we suggested that maybe the Nash equilibrium was not the more realistic of the two. But we then remarked that a deeper discussion of the point would require analysis of a dynamic game. Nothing in Sections 3.5 and 3.6 leads us to fundamentally revise our earlier conclusion. So, we are still faced with the question: did not we overstate the case against commitment? When government and private agents are thinking about the future implications of present policies, are not they likely to give the no-inflation policy a fair chance?

Secondly, a positive answer is more likely if private agents learned by experience that government is tough against temptations to inflate. The government knows it and therefore has an incentive for appearing tough, even more though than it may actually be. Understanding the learning process, and how government might want to play with it in order to generate favourable private expectations, also requires a dynamic analysis.

Both considerations concern the relation between the credibility of the anti-inflationary policy and the reputation of government. A policy maker that once inflated by surprise has, at least for a while, the reputation of not being serious in its anti-inflationary declarations. Another policy maker that has gained in the past the reputation of being tough with the temptation to sustain activity has no difficulty in making similar declarations credible. But the two considerations look at the relation from different viewpoints: the first one emphasizes interactive rationality, the second one assumes that private agents have incomplete information; the first one is made more transparent by the assumption of complete information, but has to be accurate in specifying the set of available strategies, particularly with respect to the time dimension; the second one is better understood within a simple two-period model; the first one naturally calls for inspiration from the theory of dynamic games, the second one from the theory of incentives under asymmetric information.

The two latter theories were developed on their own during the last two or three decades in order to meet the needs of other branches than macroeconomics, particularly industrial organization and the study of contracts. Both theories are already complex and difficult to master[64]. The book of Persson and Tabellini gives a fairly elaborate presentation of the implications thus far drawn from them for the analysis of government's reputation. Here we shall get just a first look at the gist of this literature and make no attempt to go into its rigorous elaboration and intricacy.

2. In order to introduce the problems of intertemporal interactive rationality in the simplest possible framework, let us consider the model of Section 3.5

[64] See, for instance, D. Kreps, *A Course in Microeconomic Theory* (Princeton University Press, 1990).

with no shocks ($\varepsilon_t = 0$) and an indefinite future, that is with an infinite horizon. Let us then consider the two following strategies, one for the private sector in forming its expectations, the other for government in deciding on the rate of inflation:

$$\begin{aligned} g_0^e = 0 \quad \text{and,} \quad \text{for } t > 0,\ g_t^e = 0 \quad \text{if } g_\tau = 0 \quad \text{for all } \tau < t, \\ g_t^e = \bar{a} \text{ otherwise,} \end{aligned} \tag{100}$$

$$\begin{aligned} g_0 = 0 \quad \text{and,} \quad \text{for } t > 0,\ g_t = 0 \quad \text{if } g_\tau^e = 0 \quad \text{for all } \tau \leqslant t, \\ g_t = \bar{a} \text{ otherwise,} \end{aligned} \tag{101}$$

where $\bar{a}$ is given by (97). Each strategy makes sense: private agents know the benefits of avoiding inflation and are so expecting government to select the commitment policy; at least such an expectation will hold as long as agents will not observe a different policy choice; but they stand ready to shift to the expectation of the equilibrium discretionary policy, if and when they will observe a deviation from the commitment policy. Similarly, government perceives the threat that private agents would be likely to expect inflation after a surprise move to it; but as long as inflation is not expected, government does not perceive existence of the credibility constraint. Moreover, the couple of the two strategies seems to make sense as a rational solution to the dilemma to which the policy choice is exposed.

More precisely strategy (101) is the best response to expectations (100), at least if the discount factor γ in (83) is not too small. Indeed, strategy (101), in conjunction with (100), will lead to $g_t^e = g_t = 0$ for all t. A surprise move to inflation in period t would lead to a higher value of Z_t but at the cost of lower values of all subsequent Z_τ. A similar argument may be formalized for private agents, for instance, if they are assumed to suffer a cost proportional to $(g_t^e - g_t)^2$. Thus we may say that the couple (100)–(101) defines a Nash equilibrium of the dynamic game.

But the argument does not imply that this Nash equilibrium would supersede the discretionary equilibrium defined in 3.5 and supposed to be applied repeatedly period after period. Both equilibria appear as valid solutions of the dynamic game, at least as long as we stick to the pure Nash's logic. And other Nash equilibria of the same game exist. We asked whether reputation can substitute for commitment in sustaining a low inflation policy. At this point the answer is: "Yes, may be".

The multiplicity of Nash equilibria reveals existence of a modelling difficulty, which generally applies to dynamic games. The theory of games indeed proposes various ways for refining the equilibrium concept, so as to eliminate a number of Nash equilibria, or for selecting one equilibrium among all of them. Persson and Tabellini discuss the conclusions to be drawn for the macroeconomic-policy problem; they stress the possible role of a concept of "sequential rationality"; but they do not find its implications fully persuasive. They suggest that complete information of private agents may be too extreme an hypothesis for our present need and that the appropriate definition of credibility may have to depend on the economic context.

3. Incomplete information is part of any full picture of the credibility problem. The reason why choice of policy measures interacts with private expectations often is that the choice may reveal something about government's true preferences: from the known measures taken by a government agents may infer that it is less or more inclined to inflate than they had formerly believed. But authorities are aware that such inference can be drawn; so their choice may be partly motivated by the purpose to later take advantage of misinterpretations of their objectives: the reputation of being tough against inflation might be gained by actions at an early stage and would then later give more effectiveness to any stimulation of activity that might be desired. But again, shrewd private agents may realize that such a motivation exists and take it into account when forming their expectations. What will then the outcome be?

These kinds of question and interaction are usually studied in models with two or several possible types of government, private agents having uncertain beliefs about the exact type of their government. Here we may have a brief look at such a model in which for simplicity there would be just two periods (1 and 2), just two types of government (tough and weak) and no shock in the economic system (Equation (90) would apply with $\varepsilon_{t+1} = 0$). Government has to decide on g_0 in period 1 and on g_1 in period 2. Before observing government decision, private agents give probability π_0 to the event that their government is weak ($0 < \pi_0 < 1$); on the basis of this probability they form their expectation g_0^e (a real number, which may be taken as a mathematical expectation). After observation of g_0 they revise their beliefs; the probability becomes π_1 (which may be equal to 0 or to 1); their expectation for period 2 is accordingly given by g_1^e. Private expectations are observed by the government (alternatively it could compute them from the model).

For simplicity again, we consider the case in which a tough government would not care at all about the level of activity. Under all circumstances it will decide $g_0 = g_1 = 0$. On the other hand, the government, if weak, has the objective function defined by $Z_0 + \gamma Z_1$ with $Z_t = (k\hat{q} - q_{t+1})^2 + b(g_t)^2$. Minimization of Z_1 under (90) leads to:

$$g_1 = \left[\alpha^2 g_1^e + b\bar{a}\right] \cdot \left[\alpha^2 + b\right]^{-1}, \tag{102}$$

where $\bar{a}$ is taken as defined by (97). The rate of inflation is equal to $\bar{a}$ if $g_1^e = \bar{a}$; it is positive but smaller than $\bar{a}$ if $0 \leqslant g_1^e < \bar{a}$. The problem for the government, if weak, is just to wisely select g_0, taking into account the impact of its choice on g_1^e.

The first temptation might be to select $g_0 = \bar{a}$, which would be the time-consistent choice in the case of complete information of private agents. But the choice of any positive value for g_0 would reveal that the government is not tough; agents, who by assumption here know α and b, would then forecast with certainty $g_1 = \bar{a}$ in period 2; so $g_1^e = \bar{a}$ would result from the choice of $g_0 = \bar{a}$.

The government will then consider taking advantage of the incomplete information of the private sector and wonder whether it ought not to rather mimick the decision of a tough government and select $g_0 = 0$. The welfare Z_0 would then be lower in period 1, because of a too low level of activity q_1; but welfare Z_1 would be higher in period 2, because less price inflation would be required, given the more favourable price expectations. What to do, of course, depends on values of the parameters.

4. In order to well understand the formation of equilibrium in such a framework we may try to characterize the roles of the prior beliefs π_0 and of the discount factor γ, whereas we may fix the values of α and b as being both equal to 1. We easily find that the following equation then holds:

$$Z_t = \left(\bar{a} - g_t + g_t^e\right)^2 + (g_t)^2. \tag{103}$$

In order to form their expectations rational private agents have to take account of the fact that a weak government might not decide $g_0 = \bar{a}$ but rather $g_0 = 0$. Let φ be the probability of this second choice by that government. Then, someone ignoring whether government is tough or weak, but knowing φ, realizes that his or her probability of observing $g_0 = 0$ is equal to $1 - \pi_0 + \pi_0\varphi$ and the probability of $g_0 = \bar{a}$ is correspondingly $(1 - \varphi)\pi_0$. The mathematical expectation of g_0 so is:

$$g_0^e = (1 - \varphi)\pi_0\bar{a}. \tag{104}$$

Assuming now that in period 1 private agents observe $g_0 = \bar{a}$, they give probability $\pi_1 = 1$ to the event that government is weak; and they expect $g_1^e = \bar{a}$.

But, if agents observe $g_0 = 0$, they are not sure about the type of their government, because with probability φ a weak government would have chosen

$g_0 = 0$. The value of the posterior probability π_1 of government to be weak, conditional on having observed $g_0 = 0$, is given by "Bayes rule":

$$\pi_1 = \frac{\varphi\pi_0}{1 - \pi_0 + \pi_0\varphi}. \tag{105}$$

The mathematical expectation g_1^e after observation of $g_0 = 0$ is equal to $\pi_1\bar{a}$, which is a function of φ through (105).

But rational agents know that a rational weak government will adapt its decision in period 1 in conformity with the strategy that will minimize its objective. The decision may then be not only to select for sure $g_0 = \bar{a}$, or for sure $g_0 = 0$, but also to select at random between $\bar{a}$ and 0, then applying the most favourable probability. The rational agents are able to compute that probability since, by assumption here, they know the characteristics of the environment in which the public strategy is adopted; so, the probability φ that they attribute to the choice of $g_0 = 0$ by a weak government is precisely that probability which is optimal from the viewpoint of the government in question.

In other words, in order to solve our problem we have to find the probability φ which will minimize the expected value of $Z_0 + \gamma Z_1$ under the conditions described thus far.

If $g_0 = \bar{a}$, which will reveal that the government is weak and therefore imply $g_1^e = g_1 = \bar{a}$, then g_0^e being given by (104), $Z_0 = [1 + (1 - \varphi)^2\pi_0^2]\bar{a}^2$ and $Z_1 = 2\bar{a}^2$. Hence $Z_0 + \gamma Z_1$ will be equal to $\bar{a}^2$ multiplied by:

$$L_R = 1 + (1 - \varphi)^2\pi_0^2 + 2\gamma. \tag{106}$$

If $g_0 = 0$, then $g_1^e = \pi_1\bar{a}$ and, in conformity with (102), $g_1 = (1 + \pi_1)\bar{a}/2$. It follows that $Z_0 + \gamma Z_1$ will be equal to $\bar{a}^2$ multiplied by:

$$L_M = \left[1 + (1 - \varphi)\pi_0\right]^2 + \frac{\gamma}{2}(1 + \pi_1)^2. \tag{107}$$

The weak government wants to select φ so as to minimize $L = \varphi L_M + (1 - \varphi)L_R$ taking into account (105), (106), (107), together with the fact that φ must belong to the interval [0, 1].

The solution of this mathematical problem is tedious and complex. It will not be given in this book, since the main reason for presenting the problem was to expose the logic of a reputation equilibrium under incomplete information. It suffices for the purpose to report that a rather large area of parameter values of π_0 and γ should, according to this logic and to this example, lead a government concerned with the level of activity to hide the fact, with probability 1, during the initial period after its appointment. The reader will find it the book of Persson and Tabellini the result obtained for another example in which the government's objective function is linear rather than quadratic. In both examples low values of the discount factor are required for the reputational incentives to be neglected by the incoming government.

Concluding on what was learned by the study of reputational equilibria of monetary policy, Persson and Tabellini express some scepticism: the literature surveyed "was more concerned with the tools of analysis than with substantive economic problems"; results are not very convincing; empirical implications are not clearly borne out by observation. Indeed, the literature addressed itself to meaningful questions, but the logic of interactive rationality was probably pushed too far to appear fully realistic, at least on the problems discussed in this chapter.

3.8. At the frontier between macroeconomics and political economy

Persson and Tabellini are less sceptical about another literature, which they also survey. It deals with the political conflicts concerning monetary policy, which often features among the stakes of elections and political debates, and which may in turn be affected by the results of elections. The new political economy also deals with political institutions, on which we touch in Section 3.6; it then takes a positive view, rather than the more normative approach in that section. However, we shall not enter here into a discussion of the political aspects of economic policies. This choice of the author requires a short explanation.

Persson and Tabellini convincingly argue that political conflicts and institutions add new strategic interactions because of the politicians motivation to please the voters. Such interactions may constrain economic policy and make it better determinate; on the other hand, elections may lead to changes in government and thus create uncertainty about future economic policy. Persson and Tabellini then show how political incentives can induce a "political business cycle", political commitments or signalling, and policy stands, alternating with the phase of the cycle. Ideological polarization can also lead to frequent changes in government and to excessive policy volatility, linked to uncertainty about the results of elections. However, elections can reinforce the reputational incentives of a government, because voters can punish the latter by voting it out of office in case of policy surprises; this enforcement mechanism can strengthen the credibility of policy announcements and bring about equilibria with low inflation.

The main reason for refraining from entering into the study of those aspects is that the interlude in this chapter is already long enough. We started the part with the recognition that we were going to examine a fiction, that of a government directly controlling the rate of inflation. We went on, realizing along the way that the fiction was perhaps not misplaced because it helped when we were discussing the formation of expectations and the credibility of policies confronted with the dilemma between high inflation and low activity. But it is now time to go back to the main thread of the chapter. We shall have another opportunity, at the end of Chapter 9, to consider the political aspects of economic policies, and then in a less narrow context than the one imposed by a fairy-tale about the monetary policy and about the true nature of the inflation–depression dilemma. A noticeable outcome of Chapter 9 will be a recommendation addressed to economic policy advisers, namely to concentrate on economic analysis and to leave political analysis to political advisers, of which there are many. With this recommendation in mind, it is all the more urgent to resume the theoretical development interrupted at the end of Part 2.

Persson and Tabellini are likely to accept the idea. They discuss at length in their book two forms of constraints on economic policies, the credibility constraints and the political constraints, both concerning the incentives to which policy makers react and are known to react. But they also mention in passing that the "knowledge constraint" on the formation of economic policy may be more binding than the incentive constraints, because policy makers do not well know how the economy will react to their interventions. Indeed! ... and this book is intended precisely to somewhat relax the knowledge constraint.

4 The Aggregate Demand Approach to Dynamic Macroeconomic Policies

In Part 2 of Chapter 7 we considered how the Keynesian system explained the effects of macroeconomic policies. We are now going to proceed to a similar study, using the dynamic model whose elements have been built in Part 2 of this chapter. Figures 2 and 4 show that the model contains two loops, involving $\{p_t\}$ and $\{d_t\}$ for the first, $\{q_t\}$ and $\{d_t\}$ for the second. This reflects the same kind of interdependence as was recognized in Chapter 7, between demand and output according to Sections 2.2 and 2.3 of that chapter, between demand and the price level according to Section 2.5. The present part can so be seen as extending the study then presented, making it both dynamic and no longer focussed on the extremes of either full price rigidity or output rigidity.

The two following sections will neglect the possibility that changes in the values of the instruments of economic policy could directly react on the expectations held by private agents. We shall then discuss the case of adaptive expectations reacting to the results of policies not to their announcement. In Section 4.3 we shall consider how alternative hypotheses about expectations, particularly rational expectations, could affect the results. Finally Sections 4.4 and 4.5 will show how the approach of this part connects with that of the foregoing part.

As in Part 2 of Chapter 7, we shall first deal, in turn, with the two cases, where the instrument of monetary policy is either the nominal interest rate r_t or the money supply M_t. The first case, discussed in Section 4.1, is obviously the simplest since it means considering a model quite similar to that of the multiplier, with just two complications. Firstly, we must follow the time path of the repercussions of policy measures. Secondly, we must consider repercussions concerning not only quantities but also prices.

4.1. Control of the interest rate with adaptive expectations

1. In this section we can ignore (75) which gives the demand for money. Indeed, we are not now particularly interested in the quantity of money. Neither are the monetary authorities, which, according to the assumptions made in this section, leave it free to adapt to whatever is required by private agents at the ruling interest rate. Discussing fiscal policy in the present framework is thus the equivalent to studying the evolution of prices and quantities (p_t and q_t) in alternative scenarios which, identical with respect to the time path of the interest rate r_t, differ in terms of the time path of the variable A_t appearing in Equation (58). Discussing monetary policy is equivalent to studying the same evolution in scenarios which include identical series $\{A_t\}$ but different series $\{r_t\}$. Clearly, these two studies are very similar: they both deal with a policy of aggregate demand management.

Formally in our analysis fiscal policy will be assumed here to act instantaneously and completely on demand, whereas the impact on demand stimulated by an action on the interest rate will extend beyond the period in which it will occur. But the difference is in part artificial, since it does not take into account the political and administrative processes through which a budgetary policy aimed at changing public consumption or investment must pass before developing into a concrete demand for goods. Moreover, we are neglecting for simplicity fiscal policy measures aimed at changing private disposable incomes; we are so neglecting the lags that would then occur between those changes in incomes and induced changes in the private demand for goods and services. These limitations, accepted for analytical convenience, should be remembered as we are going to first study the impact of a change in A_t. Afterwards, we shall easily examine how the conclusions obtained can be transferred to those concerning the impact of a change in r_t. We shall leave to the interested reader the task to draw more accurate conclusions about budgetary policies, thanks to a distinction between the time of a policy measure and the times of its impacts on the series $\{A_t\}$.

The endogenous variables of the analysis are here q_t, g_t and d_t. They are linked to each other by equations expressing the adjustment process and the formation of demand, that is, Equations (44), (45) and (58). Solving this system leads to:

$$D(L)\big(q_t - \hat{q}\big) = k_1 L(1 - cL)z_t, \tag{108}$$
$$D(L)\big(g_t - \hat{g}\big) = (k_3 - k_5 L)(1 - bL)z_t \tag{109}$$

with:

$$D(L) = (1 - cL)\big[1 - k_2 L - k_1 L A_q(L)\big] - (1 - bL)(k_3 - k_5 L)A_p(L), \tag{110}$$
$$z_t = A_t + A_q(1)\hat{q} + A_r(L)\big(r_t - \hat{g}\big). \tag{111}$$

In Equations (108) and (109) we again come across the dilemma which was studied in Section 2.2 and is due to the nature of the adjustment process. The impact on q_t is faster than the impact on p_t. This impact is highest if k_1 is high and k_3 is low, that is, if the rate of capacity utilization is low. On the other hand, the impact on prices dominates and tends to be faster (k_5 positive) when the rate of utilization is high[65].

2. If we want to go further we must take into account the expression for $D(L)$. Without any close examination we can, however, exhibit the *long-run relationship*, which would be established if $A_t + A_r(L)r_t$ were constant and equal to A, and if q_t and g_t were also constant equal to q and g (remember, however, that all along this part we are assuming, as in Part 2, an exogenous and constant productive capacity; the "long-run" is therefore only a reference for the dynamic evolutions generated by a model built for the study of the short run). These relationships are obtained directly from Equations (108) and (109), since L before a constant can be replaced by 1. We thus obtain:

$$D(1)\big(q - \hat{q}\big) = k_1(1 - c)\big[A + A_q(1)\hat{q} - A_r(1)\hat{g}\big], \tag{112}$$
$$D(1)\big(g - \hat{g}\big) = (k_3 - k_5)(1 - b)\big[A + A_q(1)\hat{q} - A_r(1)\hat{g}\big]. \tag{113}$$

They have an exceptionally simple expression when expectations are not regressive ($a = c = 1$); that is:

$$q = \hat{q}, \qquad g = \frac{(1 - a_1 - a_5\delta)\hat{q} - A}{-a_4\delta}. \tag{114}$$

[65] Taking into account the expression of A_t given in footnote 50, we see that, when $a = c = 1$, the constant $\hat{g}$ disappears from (109), since in this case $A_r(1) = A_p(1)$.

So, in this case which may be the most interesting, the only possible stationary equilibria are those in which the rate of capacity utilization is equal to its optimal value: this is a direct result of previous assumptions as to the adjustment process, as expressed by the two equations (36) and (37).

It may be more surprising to observe that, in this case, the stationary rate of inflation is shown to be a decreasing function of the level A of autonomous demand (a_4 is negative). However, the result can be explained since productive capacity is given and since in the long run the demand induced by economic activity does not depend on the rate of inflation, except through the real interest rate; but this induced demand is a decreasing function of the real interest rate, hence an increasing function of the rate of inflation: the higher autonomous demand, the lower induced demand must be for there to be equilibrium, hence the lower must the rate of inflation be. But the paradoxical nature of this property raises the question of knowing whether the stationary equilibrium found may be realistic: would it actually tend to occur or is it merely an abstract concept without empirical counterpart?

Before answering the question, which requires considering the evolution out of the stationary equilibrium, let us examine also what this equilibrium can be in the case of regressive expectations. In general, $D(1)$ is derived from (110) and given by:

$$\frac{D(1)}{1-b} = (1-a)\big[1-k_2+k_1(1-a_1-a_5\delta)+(k_3-k_5)a_2\big] + a(k_3-k_5)a_4\delta. \tag{115}$$

The term in square brackets is positive, whereas the last term is negative. If the tendency to expect a return to the normal rate of inflation $\hat{g}$ is strong enough (a is small enough), and if investment is not too sensitive to the interest rate ($-a_4$ fairly low), $D(1)$ is positive. Assuming this case occurs and seeing that the multipliers of A in (112) and (113) are also positive, we arrive at the natural result that, in the stationary equilibrium, the rate of capacity utilization as well as the rate of inflation are increasing function of autonomous demand. In this case the increase in the rate of inflation has a depressive effect on demand, an effect which is stronger than the stimulation produced by the fall in the real interest rate: on the one hand, production is stimulated and the rate of capacity utilization is maintained at a higher level; but on the other hand, consumption is depressed because the increase in the expected rate of inflation is permanently smaller than that of the rate of inflation actually occurring (the expected deceleration of inflation is maintained at a higher level).

3. To go further and examine *short-run developments*, we must study the complete dynamic process, as defined by Equations (108) to (110). We want to see in particular if this process would tend to converge in the case where z_t would remain constant after a certain date. The result depends only on the autoregressive form appearing in the left-hand term of (108) and (109), i.e. on the properties of $D(L)$. The process will converge no matter what the initial conditions are if, and only if, the roots of the polynomial in u given by the numerator $N(u)$ of the rational function $D(u)$ all have absolute values greater[66] than 1. We must, therefore, study this polynomial.

Although we tried to choose convenient specifications for behavioural laws, the study of the polynomial $N(u)$ remains complex because it is in general of the fifth degree. In fact, its properties depend on numerous parameters, which were introduced one after the other. We shall make here just a partial exploration of the properties; it will, nevertheless, be enough to show the following two conclusions:

– the initial impact of measures affecting aggregate demand can be very different from the impact reached after a number of periods;
– only precise econometric investigations can possibly establish the dynamic properties of the system generated by the behaviour of agents and the adjustment process.

Note, however from the start that, in the case where expectations are not regressive, the stationary equilibrium defined by Equations (114) is unstable; thus it does not tend to materialize; paradoxical properties of these equations are so explained. Indeed, we immediately see from (110) and (66), that $D(0) = 1$. Moreover, when $a = 1$, Equation (115) shows that $D(1)$ is negative. If u changes from 0 to 1, $D(u)$ changes continuously passing from a positive value to a negative value. Hence, there is a value u_0 of u between 0 and 1 such that $D(u_0) = 0$, therefore also $N(u_0) = 0$. So the evolution does not converge, except when it starts precisely from the values found in the stationary equilibrium.

It is clear that the same instability of the stationary equilibrium prevails so long as $D(1) < 0$, that is so long as the last term in Equation (115) dominates. We may say that *the instability of the inflationary process appears as a char-*

[66] We can easily understand why this is so if we consider the solution of $y_t = ay_{t-1} + x_t$, which can be written

$$y_t = a^t y_0 + \frac{1 - a^t L^t}{1 - aL} x_t.$$

In the case where x_t is constant, convergence will occur precisely if $|a| < 1$; and a indeed is the inverse to the root of $1 - au$. For the general theory see the mathematical digression in Chapter 9 (Section 1.2).

acteristic of those economies which are very sensitive to inflation. Moreover $D(1) < 1$ is also the necessary and sufficient condition for the stationary equilibrium to have the property that we earlier found paradoxical when looking at the implications of Equations (114). Now, the paradox disappears since the property has no chance to be actually observed.

Going further, consider the limit case in which k_1 would be nil. Pressure of demand would not influence output at all, but only prices. This is obviously an extreme case, and for this reason, unrealistic. However, it can give us an acceptable first approximation for the cases in which the economy operates at a high level of capacity utilization.

If $k_1 = 0$, then $k_5 = k_2 k_3$; the term $(1 - k_2 L)$ appears as factor in $D(L)$ and can be eliminated from (109), which is then written as:

$$\big[(1 - cL) - k_3(1 - bL)A_p(L)\big]\big(g_t - \hat{g}\big) = k_3(1 - bL)z_t. \tag{116}$$

Taking account of expression (66) of $A_p(L)$, we see that the polynomial whose roots we should study, is here:

$$\begin{aligned} N(u) \;=\; & (1 - cu)(1 - \nu u)(1 + a_2 k_3 u) \\ & + a a_4 k_3 (1 - \nu)(1 - b)(1 - \alpha u)u^2. \end{aligned} \tag{117}$$

Stability holds if expectations are sufficiently regressive $(c < 1)$ *and if the retroaction of the rate of inflation on itself through aggregate demand is weak* (if $a_2 k_3$ and $a_4 k_3$ are close to zero, then $N(u)$ has a negative root, large in absolute value, and two positive roots close to $1/c$ and $1/\nu$, respectively, and thus greater than 1). But we also see that the product of the roots is smaller than 1, and thus there *is instability, if the feedback in question is large* ($a_2 k_3$ and $-a_4 k_3$ high).

Other cases of instability of the process defined by Equations (108) to (111) can also be found. Indeed, we see that in general the product of the roots of $N(u)$ is the inverse of:

$$\gamma\big\{a_2 c \nu k_5 + \alpha(1 - \nu)\big[a_5 c k_1 - a_4 a (1 - b) k_5\big]\big\}. \tag{118}$$

If the absolute value of this term is higher than 1, instability prevails. Beside the cases we saw above, and which can hold even when k_1 is not nil, we still recognize that *instability can result from the accelerator effect* (if $k_1 a_5 (1 - \nu)$ is high, which is more likely to happen the higher the rate of capacity utilization).

4. In all cases of instability, effects after a number of periods have a good chance to much differ from short-run effects. In principle they can be computed by recursive application of Equations (108) to (110).

Suppose that, at date $t = 1$, starting from a stationary equilibrium where $z_t = z_0$, $q_t = q_0$ and $g_t = g_0$ for $t \leqslant 0$, a permanent decrease in autonomous demand is decided with the aim of slowing inflation down ($z_t = z_0 - \varepsilon$ for $t > 0$). So we calculate:

$$g_1 = g_0 - k_3\varepsilon, \qquad g_2 = g_0 - (1+\alpha_1)k_3\varepsilon, \tag{119}$$

where:

$$\alpha_1 = a(1-b) + k_2 - \frac{k_5}{k_3} - k_3 a_2 - k_1\big[1 - (1-\gamma)a_1\big]. \tag{120}$$

Beyond these two initial periods, formulas very quickly become unmanageable. It is already worth noting, however, that the decrease in g_2 may be smaller than that in g_1 because α_1 can be negative and will indeed be if a_2 is large: an unexpected decrease in inflation in period 1 would lead to high real income, hence to high consumption in the following period, which would then mitigate the decrease in autonomous demand.

As a numerical example, consider the case where $a = c = 1$, $k_1 = 0$, $k_3 = 1$, $a_2 = 1$, $a_4 = -1$, $b = \nu = 3/4$ and $\delta = 1/8$ (so $\alpha = 7/8$). Computation leads to:

$$\begin{gathered} g_1 = g_0 - \varepsilon, \qquad g_2 = g_0 - \frac{\varepsilon}{4}, \qquad g_3 = g_0 - \frac{21}{16}\varepsilon, \\ g_4 = g_0 - \frac{65}{128}\varepsilon, \qquad g_5 = g_0 - \frac{419}{256}\varepsilon. \end{gathered} \tag{121}$$

We see that, in this case of instability which is close to the borderline of stability, the decrease in autonomous demand would have a lasting effect on slowing down inflation.

5. The main interest of Equations (108) to (110) may lie in the possibility of using them in the reverse direction in order *to determine the time profile of an anti-inflationary policy* acting through aggregate demand. Fiscal measures affecting A_t or monetary measures acting on the interest rate r_t change the time path of z_t, hence also that of the rate of inflation g_t (and incidentally also that of the rate q_t of capacity utilization). In order to gear the rate of inflation down along a given path, one may like to know what should be the evolution of

z_t for the purpose. Equation (109) indicates exactly how this should be done, if we read it as determining which time path of z_t has to be chosen in order to obtain a given time path of g_t.

Let us consider the ideal case where a stationary equilibrium would have been achieved up to the present: $g_t = g_0$, $q_t = q_0$ and $z_t = z_0$ for $t \leqslant 0$. Let us assume the objective is an immediate and permanent fall in the rate of inflation, so as to reach $g_t = g_0 - \eta$ for $t > 0$. What values must be given to the sequence of the z_t?

We can recursively compute these values and obtain:

$$z_1 = z_0 - \frac{\eta}{k_3}, \qquad z_2 = z_0 - (1 - \alpha_1)\frac{\eta}{k_3}, \tag{122}$$

where α_1 is given by Equation (120) (as noted earlier α_1 can be negative, so that a larger decrease in autonomous demand may be needed in period 2 than in period 1 for achieving the same decrease in inflation). But the formulas quickly become unmanageable. Since we can assume that k_5 is lower in absolute value than k_3, since we know that b is positive and smaller than 1, and that the denominator of $D(u)$ is a product of monomials whose roots are positive and smaller than 1, we have, anyway, the guarantee that the sequence of the required z_t converges towards:

$$z_{\lim} = z_0 - \frac{D(1)\eta}{(k_3 - k_5)(1 - b)}. \tag{123}$$

This last formula is interesting because we saw that $D(1)$ can be negative and even has to be if expectations are not regressive ($a = c = 1$). If this is so, the policy of aggregate demand will be first of all restrictive, but will in the end become expansionary. The reason is that, given a fixed nominal interest rate, the lasting reduction in the rate of inflation would tend to become depressive.

To be more specific, let us confine ourselves just to monetary policy, that is let $A_t = A_0$ for all t, and let us assume $a = c = 1$. We can then directly calculate the series of the required interest rates, with in particular:

$$\begin{aligned} r_1 &= r_0 - \frac{\eta}{k_3 a_4 (1 - \nu)}, \qquad r_2 = r_0 - \frac{(1 - \alpha_1 + \alpha - \nu)\eta}{k_3 a_4 (1 - \nu)}, \\ r_{\lim} &= r_0 - \eta. \end{aligned} \tag{124}$$

The policy must start with an increase in the interest rate. But after a while the interest rate must fall to a lower value than it initially had, and this by an

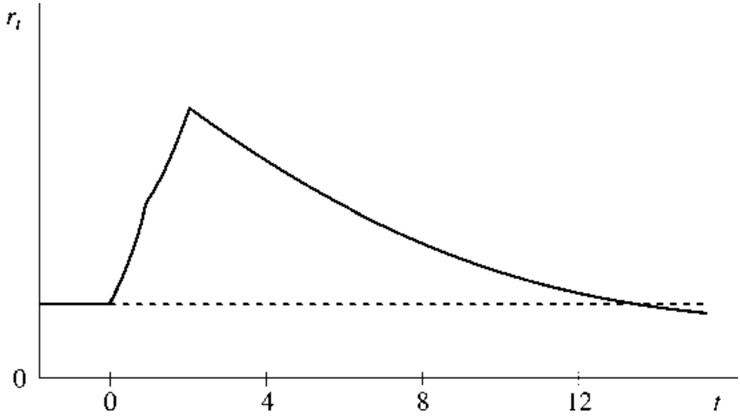

Fig. 5.

amount eventually just equal to the reduction in the rate of inflation. However, the required reversal may not happen soon. We note also that r_2 may have to be larger than r_1 for achieving the same reduction in inflation; the reasons are similar to those explaining earlier that z_2 might have to be smaller than z_1.

As an example, consider again the case in which $k_1 = 0$, $k_3 = 1$, $a_2 = 1$, $a_4 = -1$, $b = \nu = 3/4$ and $\delta = 1/8$. This really is a limit case, which systematically favours the success of a monetary policy of deflation, since changes in aggregate demand are assumed to affect only prices and not production (clearly, such is seldom the case). We compute:

$$\begin{aligned} r_1 &= r_0 + 4\eta, \\ r_t &= r_0 - \eta + \left[\frac{15}{2}\left(\frac{7}{8}\right)^{t-2} + \left(\frac{3}{4}\right)^{t-2}\right]\eta \quad \text{for } t \geqslant 2. \end{aligned} \tag{125}$$

The interest rate must thus rise sharply in the early periods. If we want to achieve a fall of 1% in the rate of inflation per unit period, we must raise the unit period interest rate by 4% in the first period, 7.5% in the second and 6.3% in the third; there is a progressive decline in this rate afterwards, but it becomes smaller than its initial value only after the seventeenth period[67] (see Figure 5).

[67] Since the unit period was chosen to be about a quarter, the policy would achieve a fall of about 4% per annum in the rate of inflation; if initially the rate of interest was 12% per annum, the policy would require an interest rate exceeding 25% per annum during six quarters. It might be argued

Actually the increase in the interest rate appears so sharp during four to six periods that monetary authorities, though determined to reduce inflation, would probably reconsider their policy choice and opt for a progressive decrease in the inflation rate rather than for an immediate decrease.

These numerical results are, however, no more than illustrative. Only econometric studies can help to precisely know the behavioural laws and their parameters, the lags in adjustments and the like. Such studies are prerequisite in order to precisely determine the strategy to adopt in order to slow down inflation through actions on the interest rate. We shall more generally come back to this subject in the following chapter and already present some results about it later in this chapter (see Sections 6.4 to 6.6).

Nevertheless, this study completes what was discussed in Part 6 of Chapter 5 concerning financial growth strategies. We saw why the least inflationary steady growth paths had the lowest nominal interest rate. But we left aside the question of knowing how an interest rate policy could allow to pass from one steady growth path to a less inflationary one. We now see, qualitatively, how this can be done.

4.2. Control of the money supply with adaptive expectations

1. Substituting the assumption of control over the money supply for that of control over the interest rate does not fundamentally alter the analysis but makes it more complex. We must no longer introduce just the adjustment process (44)–(45), and Equation (58) which characterizes the pressure of the demand for goods; we must also bring in Equation (76) which characterizes the equilibrium on the money market. But solving this system will lead again to two equations linking respectively the time paths of q_t and g_t to the time paths of the exogenous variables. The interest rate r_t will no longer figure among these, but be replaced by a variable characterizing the "money supply", that is, the rule followed by monetary policy. The new step in the theoretical approach is identical to that made in Chapter 7 where, after we had examined the theory of the multiplier, we studied the consequences of the joint realization of equilibrium on the goods market and on the money market.

The most natural assumption is to suppose that monetary authorities fix the value of the liquidity ratio m_{t+1} in each period. Equations similar to (108) to

that the forms of the equations and the values chosen for the parameters would be better suited for a longer unit period, this being for instance of a semester. If so, a fall of 2% per annum of the inflation rate would require less extreme interest rates but for a longer period.

(111) will still define the dynamic functioning of the system; there will be just one change, namely replacement of $A_q(L)$, $A_p(L)$, A_t and z_t respectively by:

$$A_q^*(L) = A_q(L) - \frac{1}{b_2} B_q(L) A_r(L), \tag{126}$$

$$A_p^*(L) = A_p(L) - \frac{1}{b_2} B_p(L) A_r(L), \tag{127}$$

$$A_t^* = A_t + \frac{1}{b_2} A_r(L)(m_{t+1} - B_0), \tag{128}$$

$$z_t^* = A_t + A_q^*(1)\hat{q} + \frac{1}{b_2} A_r(L)\bigl(m_{t+1} - B_0 - b_2\hat{g}\bigr). \tag{129}$$

Indeed Equations (108) to (111) were constructed by separating, in the pressure of demand (58), the parts which came from q_t and g_t from what came from exogenous variables. If we eliminate r_t using (76) and (58), we end up with a new expression for d_t which exactly leads to the replacement specified above.

The first remark about these changes concerns the initial autonomous impact on the demand for goods: in z_t the interest rate r_t is replaced by m_{t+1}/b_2; the more elastic the demand for money with respect to the interest rate the weaker will this impact be. As in Chapter 7 the possibility of a "liquidity trap" appears.

The second remark is to observe that the dynamic properties of the system are only affected through $D(L)$. In particular $D(1)$, given by Equation (115), should be replaced by $D^*(1)$ given by:

$$\frac{D^*(1) - D(1)}{1-b} = \bigl[(1-a)k_1 b_1 + a(k_3 - k_5)b_3\bigr]\frac{a_4\delta}{b_2}. \tag{130}$$

Since b_2, b_3 and a_4 are negative, $D^*(1)$ will be even more negative than $D(1)$ if expectations are not very regressive (a close to 1) and if the demand for money is sensitive to the rate of inflation (b_3 large in absolute value). Instability will tend to be reinforced. On the other hand, the dynamic properties will be close to those obtained in the case of a controlled interest rate if the demand for goods is not very sensitive to the interest rate: if a_4 is low in absolute value, then $A_r(L)$ is small and $D^*(L)$ little differs from $D(L)$.

2. So the problems which arise from choosing a monetary strategy concerning the liquidity ratio are not very different from those when choosing a strategy concerning the interest rate. But the model introduced here allows us to consider another form of strategy, which has been attracting a lot of atten-

tion since it was recommended by M. Friedman[68], namely *fixing the growth rate of the quantity of money*. We should now study this. (The strategy will be considered again later, particularly when we shall refer to political economy arguments in its favour. What follows concerns only the consequences of the strategy in the short-run framework of this chapter.)

The strategy in question is very simple to formulate since it is expressed by:

$$M_{t+1} = (1 + g_M)M_t, \tag{131}$$

where the growth rate g_M is fixed. The idea is to keep a value of g_M which corresponds to the "optimal" rate of price increase, or more precisely, a value exactly equal to the growth rate which the value of aggregate output would have if full use of productive capacity were continually achieved and if prices rose at an optimal rate. The expectation is that such a rule for money emission, if maintained persistently without fail, will result after a time in prices growing at the desired rate.

We shall not, here, examine the difficulties of putting this rule into practice. There is the question of knowing which indicator to choose for the quantity of money M_t, also the question of knowing whether corrections ought not to be introduced to take account of possible institutional changes affecting the "monetization" of the economy, even whether such changes are not likely to be induced by restriction in the money supply.

On the other hand, the approach in this section is perfectly adequate to see whether, by its effects on the pressure of demand, the rule (131) for money emission will impose convergence of the rate of inflation towards the desired value. It is indeed a natural idea to introduce the value m_{t+1} of the liquidity ratio resulting from (131) into the system that determines the time path of g_t; conditions for convergence of this time path can then be studied.

Under the assumptions made here, where output at full capacity utilization is constant, Equation (131) implies:

$$m_{t+1} = \frac{1 + g_M}{1 + g_{t-1}} m_t. \tag{132}$$

If the liquidity ratio converges towards a limit value, then the rate of inflation converges towards g_M, as would be expected[69].

[68] See, for instance, M. Friedman, *A Program for Monetary Stability* (Fordham University Press, New York, 1960).

[69] The two following parts of the section concern the analytics of the monetary strategy now discussed. The reader who would not be interested by this analytics can directly go to the last paragraphs of the section, from number 5 on.

3. However, Equation (132) is the source of a serious analytical difficulty: it implies a non-linear dependence of m_{t+1} with respect to the sequence of previous values of g_t. There is no question here of introducing, into a system which is already complex, this dependence without previously linearizing it. Studying the convergence of the linearized system cannot lead to a rigorous proof of the conditions for convergence of the initial system. But in our exploration which does not claim to be complete and makes other simplifications, the study should bring out the nature, if not the exact form, of the conditions in question. It will indeed suggest that, if the chosen rate g_M is low enough to restrict the monetary emission, convergence is not certain but only likely in normal looking cases.

To define the linearized system, the most convenient reference is to an initial stationary growth with a constant rate of inflation g_0, a constant liquidity ratio m_0, a constant autonomous demand A_0, etc. (the question as to whether this growth might have been unstable has no significance since we say nothing about how it was controlled). So, we shall substitute for (132) the approximation:

$$m_{t+1} = (1 + g_M - g_0)m_t - (g_t - g_0)\overline{m}, \tag{133}$$

where $\overline{m}$ is constant; and we shall limit attention to the case where $g_M < g_0$.

Assuming that, with the system resulting from (133), convergence takes place, the limit of the rate of inflation g_t will not be exactly g_M unless the liquidity ratio m_t converges towards $\overline{m}$. This is a fault in the linear approximation substituted for (132). We need not pay attention to it.

The analytical procedure for treating the question at issue is clear from now on. The relevant system is derived from that made of Equations (126) to (129) if we replace m_{t+1} by its value resulting from (133), namely:

$$m_{t+1} = \frac{\overline{m}L}{1 - \omega L}(g_t - g_0) \tag{134}$$

with:

$$\omega = 1 + g_M - g_0. \tag{135}$$

Equations (126), (127) and (129) are then replaced by:

$$\overline{A}_q(L) = A_q^*(L), \tag{136}$$

$$\overline{A}_p(L) = A_p^*(L) - \frac{\overline{m}A_r(L)L}{b_2(1 - \omega L)}, \tag{137}$$

$$\bar{z} = z_0^* + \frac{A_r(1)}{b_2}\left[\frac{\overline{m}g_0}{g_0 - g_M} - m_0\right]. \tag{138}$$

The time path of the rate of inflation is, consequently, determined by:

$$\overline{D}(L)\big(g_t - \hat{g}\big) = (k_3 - k_5)(1 - b)\bar{z}, \tag{139}$$

where the function $\overline{D}(L)$ is determined from $\overline{A}_q(L)$ and $\overline{A}_p(L)$ by the equation directly transposing (110).

4. One indication already appears to suggest that the rule (133) helps to stabilize the time path of the rate of inflation: so long as the variable u is lower then 1 in absolute value, $\overline{D}(u)$ is higher than $D^*(u)$, (the latter being deduced from Equations (126) to (129) by transposition of (110) again). Indeed, $A_r(u) < 0, b_2 < 0$ and $\omega < 1$ result in $\overline{A}_p(u) < A_p^*(u)$. Since $\overline{D}(0) = D^*(0) = 1$, the domain of the values of the parameters for which $\overline{D}(u)$ has one or more roots in the interval $[-1, 1]$ is smaller than the domain of the values for which $D^*(u)$ has roots in that interval.

It is interesting to consider the possible limit value $\bar{g}$ of g_t. It is enough to do this for the case where expectations are not regressive, which appears to conform most closely to the ideas of those who favour the monetary rule (131). So choosing $a = c = 1$ and knowing that $\hat{g}$ is thus eliminated from Equation (139), we know that $\bar{g}$ must satisfy:

$$\overline{D}(1)\bar{g} = (k_3 - k_5)(1 - b)\bar{z} \tag{140}$$

or:

$$\overline{A}_p(1)\bar{g} = -\bar{z}, \tag{141}$$

whereas g_0 satisfies:

$$A_p^*(1)g_0 = -z_0^*. \tag{142}$$

Taking (137) and (138) into account, we find:

$$A_p^*(1)(g_0 - \bar{g}) = \frac{A_r(1)}{b_2(g_0 - g_M)}\big[\overline{m}(g_0 - \bar{g}) - m_0(g_0 - g_M)\big]. \tag{143}$$

Equations (124), (72), (73) and (78) show that the previous equality can be simplified to:

$$\frac{m_0}{g_0 - \bar{g}} = \frac{\overline{m}}{g_0 - g_M} + b_2 + b_3. \tag{144}$$

For $\bar{g} = g_M$, as (132) would require, the limit $\overline{m}$ ought to be:

$$\overline{m} = m_0 - (b_2 + b_3)(g_0 - g_M). \tag{145}$$

We note that this equation corresponds exactly to what we would have been led to expect from the demand for money (75). *Reducing the rate of inflation has the result that the limit liquidity ratio increases* (b_2 and b_3 are negative).

We would be taken too far off by making even a preliminary and partial study of convergence. We would find again cases of instability where there is a strong feedback of inflation on itself (a_2k_3 and $-a_4k_3$ high). Here is just a case where convergence occurs.

Assume that $a = c = 1$, $a_2 = 0$, $b = \nu$ and $k_1 = 0$. Calculations, which we shall not repeat here, show that, after dividing by the expression $1 - k_2u$ whose root lies outside $[-1, +1]$, the numerator $N(u)$ of the rational function $D(u)$ can be written as:

$$N(u) = R(u) + \frac{k_3a_4(1-b)}{b_2}S(u) \tag{146}$$

with:

$$R(u) = (1-u)(1-bu)(1-\omega u), \tag{147}$$

$$S(u) = u(1-\alpha u)\big[(1-b)(b_3+b_2u)(1-\omega u) + \overline{m}(1-bu)\big]. \tag{148}$$

The three roots of $R(u)$ are obvious. If we choose k_3a_4/b_2 to be fairly small, the only root of $N(u)$ which could cause instability is that in the neighbourhood of $u = 1$; but taking account of (145), we see that:

$$R(1) = (1-\alpha)(1-b)m_0 > 0 \tag{149}$$

which results in $N(1) > 0$ and the root located in the neighbourhood of $u = 1$ is in fact greater than 1. Thus, this is a case of stability.

5. We now have to understand in what way the reduction of inflation brought about by control of the money supply differs from that brought about by control of the interest rate, the policy then applying the principles laid down at

the end of the previous section[70]. We can easily imagine that the differences are not fundamental: no matter what the control assigned to monetary policy, the transmission mechanism remains the same, namely the effect of monetary discipline on the pressure of demand for goods and services.

Nevertheless, when we pass from a constant growth rate of the money supply to another rate, also constant but lower, the tension on the money market appears only progressively as the gap increases between the former trend of the money supply and its new trend. The interest rate is therefore likely to also increase progressively, at least in the first phase of the new process. So we may expect that the effect on demand and inflation is slower than if the decision had been taken to immediately raise the interest rate.

Let us take another look at an example we dealt with at the end of the previous section. Assume that, from a stationary inflation at the rate of 2% per period (say 2% by quarter) we divide the growth rate of the money supply by two and make it 1% by period; in the limit regime this will be the same deceleration as was to be immediately achieved by an interest rate policy in the example at the end of Section 4.1. Take again the favourable case where changes in the pressure of demand affect only prices and not the rate of capacity utilization ($k_1 = 0$); consider the same following parameter values as in Section 4.1: $k_3 = 1$, $a = c = 1$, $a_2 = 1$, $a_4 = -1$, $b = \nu = 3/4$ and $\delta = 1/8$; finally let us choose the following values for the parameters of the demand for money: $m_0 = 2$ and $b_2 = b_3 = -1$.

Keeping expression (132) itself, and not its linear approximation (133), we can compute recursively the values of m_{t+1}, g_t and r_t. We find in particular that, starting from $m_1 = m_0 = 2.000$, $g_0 = 2\%$, $r_0 = 3\%$, the following values characterize the first period: $m_2 = 1.980$, $g_1 = 1.75\%$, $r_1 = 5.0\%$. During the four following quarters m_{t+1} and g_t decrease, whereas r_t increases, up to $m_6 = 1.946$, $g_5 = 1.03\%$, $r_5 = 8.8\%$. This first phase corresponds to that of Figure 5 for the interest rate, although it is somewhat delayed and does not go as high (with $r_0 = 3\%$ we had found $r_2 = 10.5\%$). Later on, the rate of inflation oscillates around 1% per quarter, whereas the liquidity ratio and interest rate hardly change during a few quarters before slowly reversing their trends. Stationary values, if they were eventually reached, would correspond to $m = 2.020$, $g = 1.0\%$, $r = 2.0\%$. But the process is dynamically instable, as shown by the fact that $D^*(1)$ has a (very small) negative value; indeed oscillations appear more and more manifest after the tenth quarter.

[70] We shall deal later with the respective advantages and disadvantages of the different stabilization strategies. Here we just want to compare the time paths implied for inflation by the two strategies.

Except for noting that the speed of the effect of the monetary policy on inflation is due, in particular, to the favourable and extreme assumption $k_1 = 0$, we are not going to dwell longer on these results, which have little more value than that of exhibiting possible consequences of the model used in this section. They do suggest, however, that the simple monetary rule (131) may require a long lag before leading to its target, a point recognized by M. Friedman[71]. It might even, after a time, generate an instability which would have to be remedied. All in all, the rule may have more complex and uncertain implications than is often claimed when it is presented.

4.3. Monetary policy with alternative hypotheses on expectations

The exploration made in the last two foregoing sections provides a benchmark for further dynamic analysis of demand management policies. An important part of the required investigations, to be reported in this book and pursued in the future by macroeconomists, has to be econometric, so as to lead to better characterizations of the laws ruling the various components of aggregate excess demand and the market adjustments represented so far by Equations (44) and (45). More will be said on econometric findings later in this chapter and in the following one. Another part of the investigations to be made with respect to the bench-mark will be taken up now and concern the discussion of the effects that would have other hypotheses than the one made in the previous sections about the formation of expectations entertained by private agents. The discussion will be much more heuristic and deal with two possible revisions of the hypothesis in question. For simplicity we shall speak just of a monetary policy controlling the interest rate; if he or she so wishes, the reader will be able to transpose the discussion to a policy acting on the quantity of money or even to a budgetary policy; the formal apparatus of Sections 4.1 and 4.2 shows how transposition can proceed.

1. Let us first be clear on *what exactly is the hypothesis under discussion*. We note that just one variable g_t^e explicitly represents the beliefs that, in period t, private agents entertain about the future evolution. This is obviously restrictive in two ways: for no other variable expectations seem to matter; behaviour in period t is assumed to depend on the price level expected to materialize in period $t+1$, but not on the price level expected for more remote future periods.

[71] In The role of monetary policy, *American Economic Review* (March 1968), M. Friedman estimated that the limit regime would be about achieved after "say, a couple of decades" (p. 11).

We should not interpret these two restrictions too strictly. The second in particular does not mean that agents are in period t indifferent to what will happen to prices beyond date $t+2$, but rather that the further evolution of prices is represented well enough for the agents by the single variable g_t^e. They do not think in terms of a wide variety of future inflation profiles, but rather in terms of something like a one-dimensional family of profiles; we may imagine that these profiles roughly correspond to various degrees of acceleration or deceleration in inflation. This interpretation tells us how we should view the presence of g_t^e in the consumption function (59) and in the investment function through (60) and (61).

Actually, once the specific adaptive formula (41) has been chosen for the formation of expectations, this notion of thinking in terms of a one-dimensional family of profiles is fairly innocuous. As we saw in Sections 1.4 and 1.5, formula (41) is natural if agents believe that Equation (9) applies, with the white noise ε_t, to the process by which the price level is determined. If so, the expectations naturally entertained in period t for the rate of inflation in the future period $t+\theta$ can be determined recursively from the past evolution of g_t and from the expectations in period t for the periods $t, t+1, \ldots, t+\theta-1$. Formula (41) will again apply after appropriate replacements of $g_{t+\theta-1}, \ldots, g_t$ in the right-hand member by the expectations entertained for them. From t to $t+1$ the expectation for period $t+\theta$ will be revised depending on the realization of ε_t. In any case assuming (41) amounts to assuming also that at date t the past evolution of g_t contains all the information available for inflation in future periods, a point we shall discuss below.

Among the variables about which expectations ought to matter, the future interest rates would seem to be particularly significant in our present framework (future degrees of capacity utilization would also be significant, but will not be discussed at this point where attention focusses on demand-pull inflation). At date t the market gives the interest rate for period t, but already then behaviour may depend on expectations concerning the interest rates that will apply in subsequent periods. The same argument as above might be made, which would say that those expectations are directly function of the current rate r_t: a high current rate would lead to high values for the expectations of rates in future periods. Again, we should view the investment function of the model as encompassing not only the direct effect of changes in the current interest but also the indirect effect through the induced changes in the rates expected for future periods[72]. For instance, the desired capital-labour mix in period t would affect K_t^* and be a function of the interest rates expected up to

[72] Alternatively we could of course distinguish the short- and long-term interest rates holding in period t and argue that the changes in the variable r_t are meant to represent changes in both rates.

full scrapping of period t's investment. Actually when reading Equation (60) for K_t^* we naturally think of the case in which all these expected future rates would be equal to r_t. Such a case of static expectations would have good reason to apply if firms would believe that the interest rate followed a random walk (which would mean a transposition of Equation (26) to the process of r_t with a white noise γ_t).

2. We may wonder at this point about the reason for the asymmetry between r_t and g_{t-1}^e in Equation (61) defining *the expected real interest rate*: if ρ_t^e is meant to be representative of the expected real rates for not only period t but also a number of subsequent periods, why should not we accept the idea of adaptive expectations for the future values of the nominal interest rate, in the same way as we do for the inflation rate? This would, of course, somewhat affect the dynamic implications of monetary policy.

In the same spirit as for the rate of inflation we might indeed consider the following formula for r_t^e, viewed as a representative of the whole relevant set of present and future expected interest rates:

$$r_t^e = h_r r_t + (1 - h_r)\left[(1 - a_r)\hat{r} + \frac{a_r(1 - b_r)}{1 - b_r L} r_t\right], \tag{150}$$

where h_r would be the weight of the direct impact of the current rate[73], $1 - a_r$ the weight of a "normal rate" $\hat{r}$ in the impact of future expected rates and b_r the parameter characterizing the sluggishness in the adaptation of interest rate expectations. The reader may imagine how substitution of this value r_t^e for r_t in (61) would react on the developments presented in Section 4.1. He or she will not be surprised to learn that, for instance, the policy aiming at reducing the inflation rate by 1% per period for the future from $t = 1$ would require a still sharper initial rise in the nominal interest rate than was established in that section. As an example, take the same values of the parameters as for the example of Figure 5 but add $h_r = 1/3$, $a_r = 0.9$, $b_r = 2/3$, $\hat{r} - r_0 = -1\%$. Computation shows that, in order to engineer the profile of Figure 5, but now applying to $r_t^e - r_0$, we need: $r_1 - r_0 = 7.6$ instead of 4.0; $r_2 - r_0 = 12.3$ instead of 7.5; $r_3 - r_0 = 7.5$ instead of 6.3; $r_4 - r_0 = 5.5$ instead of 5.2.

This would not even be substantially different if we would stick to the expectations theory of the term structure of interest rates (see Section 1.7).

[73] The direct impact actually contains more than what is coming from the change in the desired input-mix; it concerns inventories building or depletion, as well as changes in the speed with which investment projects are realized.

3. More important than the revision just sketched are those that might result from the following consideration. The adaptive expectations assumed up to now do not appear realistic if we want to discuss the implementation of *a major policy change*, publicly announced and believed when it is initiated. Informed private agents are then unlikely to react as they were earlier doing, that is, merely *ex post* to the observed impact of the policy on market indicators. There seems to be an element of inconsistency in pretending to study the exact form to be given to the policy measures but in then neglecting the possibility that private agents could be making similar reckonings in order to find out what the policy could imply for them and to shape their expectations and behaviour accordingly.

It is precisely in order to appreciate the importance of this inconsistency that *we may want to assume rational expectations*. So doing is, of course, opting for the opposite extreme to adaptive expectations, an extreme that is not very realistic either since it assumes all the agents to be fully informed and fully confident in what authorities announced; it assumes also perfect ability of agents to model the situation and to correctly draw the logical consequences; these assumptions conflict with observations to which we alluded earlier. But deriving the implications of the extreme hypothesis should permit us to appreciate the robustness of results derived in the foregoing sections.

However, we realized in Section 1.5 and in Part 3 that the circularity introduced by the hypothesis of rational expectations much complicates in general the analytical treatment. So much so that we must find other ways to proceed than to squarely face the system obtained by insertion of the hypothesis into the model of the previous sections. We shall instead make two heuristic explorations, which ought to give us some intuitive feeling about the likely consequences of rational expectations. We shall first look at the behaviour of agents realizing that monetary policy is going to severely raise interest rates for a while so as to cut aggregate demand and to slow down inflation. Our purpose will be to see whether substitution of forward looking expectations to adaptive expectations makes policy appear easier or more difficult. We shall later reflect on a very simplified model so as to speculate about the consequences of the circularity introduced by the rational expectations hypothesis. Finally we shall refer in the next part of the chapter to the debate on "policy ineffectiveness", which opposed groups of macroeconomists twenty years ago.

4. Let us *assume an economy* where the physical conditions of activity would be stationary, where prices would increase at the rate of 8 per cent per annum and the annual interest rate would be equal to 12 per cent. The government of this economy credibly announces that decision was taken to lower

quickly and permanently the inflation rate to 4 per cent per annum. For obtaining this result the annual interest rate is raised to 20 per cent and will be maintained at the new level as long as required, a time span which authorities expect to be of one year; afterwards, the interest rate will be progressively reduced. How will behaviour of the private sector react?

Clearly the reaction will be faster than with adaptive expectations, although some inertia will remain. More precisely, private agents are assumed to take seriously the announcement of the policy change; they believe in the permanent decrease in the inflation rate; they also believe in a few years of very high interest rates, followed by years of progressively lower interest rates. But they are also committed in a number of ways: few if any of the investments under construction will be abandoned; contracts will be executed, with in particular their escalation clauses referring to past inflation; the real interest in the standing service of past loans will rise, which will have unexpected redistributive effects, and so on.

Some effects will not differ much from what they would have been with adaptive expectations, in particular those resulting from earlier commitments, but also many of those induced by the increase in the real cost of capital. The main differences will come from the expectation that the rise in interest rates will be temporary. This will lead to postpone some of the demands for goods that would have otherwise be maintained. The phenomenon is likely to be stronger than what would have been observed even with regressive adaptive expectations concerning interest rates.

Thinking about these differences we are led to the conclusion that, at least during the first year after the announcement of the policy, the decrease in the demand for goods and services will indeed be sharper and stronger than with adaptive expectations. This will permit government to reach the same results in terms of inflation with a less violent action on interest rates. But, and this also is noteworthy, the qualitative conclusions of the analysis of Section 4.1 do not seem to be overturned. Forward looking expectations seem to merely contribute to effectiveness of the monetary policy against inflation. Correlatively a depressive effect on output should occur.

At least such is the result of our heuristic discussion of the case in which agents would be better informed about the future than if their expectations were adaptive. A question remains as to whether a strict rational expectations theory would support the argument. We are so inclined to push the investigation further.

5. *The phenomena just described could obviously be formalized*. With reference to the framework built in Part 2 of this chapter, the differences would concern the formation of expectations and the formation of aggregate demand.

The same kind of adjustment laws as (33) and (34) would be maintained, since they capture how the current expectations, pressure of demand and degree of capacity utilization determine changes in output and prices, mainly because of adjustment costs, broadly understood. But, in view of applications such as the one considered above, the formalization ought to distinguish at least short-term and long-term expectations.

Let g_t^{se} and g_t^{le} be representative variables for respectively the two such expectations entertained in period t about future inflation rates, and similarly r_t^{se} and r_t^{le} about future interest rates[74]. Rational expectations would imply that each one of these four variables is representative of mathematical expectations such as:

$$g_t^{\tau e} = E(g_{t+\tau}/J_t) \tag{151}$$

conditional on the set J_t of informations available in period t; the variable g_t^{se} could for instance represent those corresponding to $\tau = 1, 2, 3, 4$ whereas g_t^{le} would be representative of all other $g_t^{\tau e}$.

Similarly, the formation of the pressure of demand can take into account not only the rational expectation hypothesis but also as many expectational variables as the application may require. This will mean more terms in an equation such as (58). For instance if, for simplicity here, we consider an application in which an exogenous use of capacity would be maintained and the policy choice would concern just the size of an otherwise defined profile of future interest rates, so that short- and long-term expectations need not be explicitly distinguished, we could replace (58) by:

$$d_t = A_r(L)r_t + A_p(L)g_t + A_{er}(F)r_t^e + A_{ep}(F)g_t^e + A_t, \tag{152}$$

where F would be the forward operator defined by:

$$x_{t+\tau} = F^\tau x_t. \tag{153}$$

The model would then be defined by the adjustment laws, expectational hypotheses such as (151), and Equation (152) in which the A functions would have to be specified.

[74] It does not matter for the argument here whether or not financial markets give reliable estimates of expectations about future interest rates, as is argued by the expectations theory of the term structure of market rates.

6. Studying such a model with some degree of generality would be quite tiring, in particular because of the circularity introduced by the rational expectations hypothesis. So let us consider *a deliberately simple example*, the purpose being to exactly see how the model operates and what it may imply (a general treatment runs the risk of overlooking some important aspects of the actual problem). In this spirit we make four heroic hypotheses. (It is left to the reader as an exercise to relax any one of these hypotheses.)

First, we rule out uncertainty. So in particular rational expectations are assumed to be exact expectations. (The hypothesis is not innocuous, as the reader may have realized in Section 1.5. We shall say more about it in Section 4.2 of Chapter 9.)

Second, we give a simple form to the policy whose effects will be examined. The policy controls the interest rate. After a long period of stationary inflation at the rate g_0^* with the interest rate r_0, the government unexpectedly but credibly announces a new policy: starting in period 1 the rate r_t will be given by:

$$r_t - r_M = (r_1 - r_M)\omega^{t-1}, \tag{154}$$

where ω is a positive number smaller than 1. The limit rate r_M is assumed to be smaller than r_0, but the initial rate r_1 is larger than r_0; the time profile of the interest rate is therefore similar to those discussed earlier, in particular in Section 4.1. There are three parameters ω, r_M and r_1; but it does not much matter because we do not intend to enter into a precise comparative study of consequences following from the choice of these parameters (our purpose is rather to analyse the logical consequences of the strict rational expectations hypothesis). Thus, we do not have to bother about the possibility that parameters of behaviour could depend on ω, r_M and r_1. We therefore do not distinguish between short-term and long-term expectations.

Third, we assume that the price level evolves according to the simple equation:

$$g_{t-1} = g_{t-1}^e + kd_t. \tag{155}$$

This means that we are maintaining the hypothesis of a constant degree of capacity utilization (the case in which $k_4 = 0$ in Equation (37), where k_3 is replaced by k for simplicity). There is moreover a difference in timing with Section 2.1, where a lag of one period between the pressure of demand and the impact on prices was assumed (g_{t-1} is determined at the same time as p_t). Such a lag is easily explained by price and wage adjustment costs, but it would be a bit confusing now in the analytical treatment; we therefore prefer to assume simultaneity in the determination of p_t and d_t.

Fourth, we want to assume a very simple expression for Equation (152) specifying the formation of the pressure of demand. But we want to capture two effects of changes in inflation and interest rates, as we did in Section 2.4 with a more complex specification. Unexpected changes in the purchasing power of current incomes, transfers and standing credits have a depressive (resp. stimulating) effect when the price level unexpectedly increases (resp. decreases). Also, current changes and expected changes in the real interest rates lead to changes in investments and inventories. Taking these two effects into account leads us to assume:

$$d_t = \lambda\bigl(g_{t-1} - g_{t-1}^e\bigr) + \mu(\rho_t - \rho_0) + \nu\bigl(\rho_{t+1}^e - \rho_0\bigr), \tag{156}$$

where the real interest rate ρ_t is defined by (31) and $\rho_0 = r_0 - g_0^*$. The three parameters λ, μ and ν are negative. We also assume ν being larger in absolute value than μ, because the last term in (156) is meant to be representative of expectations entertained in period t about the real interest rate in all future periods; the weight of this term in the determination of demand during period t should therefore be high, notwithstanding the fact that the prospect of an increase in the real interest rate may lead to acceleration in the implementation of some projects during period t.

In conjunction with (155) the hypothesis of exact expectations has a very strong implication, namely that the pressure of demand d_t has to be equal to zero in all periods, except the first one, when an unexpected change in the price level $g_0 - g_0^*$ may occur (the implication would be mitigated but not removed if our specification had recognized lagged effects of price surprises on demand). This type of implication is not new for us, since we already saw it in Part 3, there on the degree of capacity utilization and in conjunction with Equation (81).

For all future periods, starting with $t = 2$, evolution of the rate of inflation follows a simple process given by equality to zero of the sum of the two last terms in the right-hand member of (156), exact expectations and Equation (154) defining the profile of the policy. The equation of the process is found to be:

$$\nu(g_t - g_M) + \mu(g_{t-1} - g_M) = (\nu\omega + \mu)(r_1 - r_M)\omega^{t-1}, \tag{157}$$

where g_M is given by:

$$g_M = r_M - \rho_0. \tag{158}$$

Before going any further, let us note that our problem does not fit exactly into the framework commonly found in the literature on rational expectation models, in which the equations ruling the processes are supposed to be time-invariant. Equation (157) is time invariant, but applies only from period 2 on. Our problem is to find the forward looking solutions starting in period 1, when the process is temporarily ruled by a different equation. However, we may begin with finding what Equation (157) implies.

At first sight, this equation looks nice. It suggests convergence of the inflation rate g_t to the limit g_M, which will eventually lead to a real interest rate equal to ρ_0. From its earlier value g_0^* the inflation rate will eventually decrease by exactly the difference $r_0 - r_M$ applying to values of the nominal interest rate. Closer examination, however, reveals two possible difficulties.

Indeed, the solution of (157) written for all t larger than 1, is:

$$g_t - g_M = (r_1 - r_M)\omega^t + \left[(g_1 - g_M) - \omega(r_1 - r_M)\right]\left(\frac{-\mu}{\nu}\right)^{t-1}. \tag{159}$$

Since ω and μ/ν are both positive and smaller than 1, convergence of g_t to g_M holds. But the equation determines g_t for $t > 1$ as a function of the inflation rate g_1 observed in period 2, but exactly expected in period 1. Of course, we still have to look at the condition for equilibrium in period 1; but this condition is better viewed as determining p_1 and g_0 as a function of the parameters and of g_1. One degree of freedom therefore remains. Such cases of indetermination often occur in the dynamic study of rational expectation processes (we indeed already found a case of that sort in Chapter 6, Section 5.4). Their existence is, of course, disturbing. For the time being, let us consider g_1 as a free parameter.

But another unpleasant feature appears in (159), namely the alternating sign of the factor multiplying the square bracket. This does not seem realistic if the square bracket is not small. The idea then comes to mind of focussing attention on the case in which g_1 would be such as to give a zero value to the square bracket (the idea would even have looked imperative if we had not assumed ν larger than μ in absolute value). If so, $g_1 - g_M$ would be equal to $\omega(r_1 - r_M)$; the real interest rate following from (31) and (154) would return to the value ρ_0 already in period 2 and would later remain at that level since $g_t - g_M$ would be exactly equal to $\omega(r_t - r_M)$. But our previous discussion of anti-inflationary policy led us to think of the quantity $\omega(r_1 - r_M)$ as being large, much larger than $g_0^* - g_M = r_0 - r_M$. Having now to view $g_1 - g_0^*$ as being positive and large does not square with the normal notion of what such a policy has to early achieve in order to be credible.

We may therefore think of giving up the idea of a zero square bracket in (159). Since g_1 will then be smaller than would be necessary for that, the square bracket will be negative. For all even values of t starting with $t = 2$, $g_t - g_M$ will have to be larger than $\omega(r_t - r_M)$, which remains disturbing for the same reason as was given just earlier about g_1. We so have to realize that, in order to come as close as possible to what common sense suggests, we shall have to accept a solution giving fairly high values to g_1, g_2 and, as we shall presently see, g_0.

At this point we can hardly avoid wondering whether the hypothesis of rational expectations and the private behaviour it asserts are not too extreme to avoid paradoxical implications. (Remember that the hypothesis does not only mean that behaviour is purely forward looking; it also assumes that private agents perfectly know how the economy functions and that they base their expectations on exact solutions of the model describing this functioning.) The same uneasy feeling would hold, although perhaps less clearly, if we had introduced in (156) not a single but several terms concerning real interest rates in future periods. Faced with such paradoxical implications, we are inclined to think that some dependence with respect to past evolution has to appear in any realistic theory of the phenomena here discussed[75]. We shall come back to the difficulty at the end of the section.

But in order to be complete we should not stop at this point the study of our example. We still have to see how is determined g_0, the inflation rate in the first period of application of the unexpected policy. The joint determination of g_0 and d_1 is given by (155) and (156), the latter being written as:

$$\begin{aligned} d_1 = {} & (\lambda - \mu)\big(g_0 - g_0^*\big) + \mu(r_1 - r_0) \\ & - \nu\big[g_1 - g_M - \omega(r_1 - r_M)\big]. \end{aligned} \tag{160}$$

Indeed, (158) shows that the square bracket here is equal to $-(\rho_2^e - \rho_0)$.

As is usual with the determination of the price on a market, a stability condition is normally required in order to make the model meaningful and to avoid

[75] The reader may wonder whether drawing such far-reaching implications from a special example is legitimate. But the example was not chosen on purpose and it does not perform so badly with adaptive expectations. We may indeed add to Equations (154) and (156) an adaptive hypothesis borrowed for g_t^e from the previous sections, moreover writing $r_{t+1}^e = r_t$. With reasonable values of the parameters the evolution generated by the solution appears satisfactory. For instance, with Equation (6) and $b = 3/4$, with $r_0 = 0.03$, $r_1 = 0.10$, $r_M = 0.02$, $\omega = 7/8$, $g_0^* = 0.02$, $k = 1$, $\lambda = -0.06$, $\mu = -0.01$, $\nu = -0.04$, we find in per cent per period: $g_0 = 1.66$, $g_1 = 1.62$, $g_2 = 1.59, \ldots, g_5 = 1.51, \ldots, g_8 = 1.46$ with then $r_8 = 5.59$.

a cobweb in the mutual adaptations of price and excess demand. Here the condition writes:

$$k|\lambda - \mu| < 1. \tag{161}$$

Solution of (155) and (160) directly leads to:

$$g_0 = g_0^* + \varphi\{\mu(r_1 - r_0) - \nu[g_1 - g_M - \omega(r_1 - r_M)]\} \tag{162}$$

with φ given by:

$$\varphi = \left[\frac{1}{k} - \lambda + \mu\right]^{-1}. \tag{163}$$

This new parameter is obviously positive if $-\lambda + \mu$ is positive; it is also positive in the opposite case because (161) is assumed to hold. If ν is negative, the growth rate g_0 of the price level in period 1 is a decreasing function of g_1: choosing a lower g_1 implies a higher g_0.

Equations (159), applying for $t > 1$, and (162) determine the whole sequence of inflation rates, with however one degree of freedom, as we already realized. The conclusion of our exercise is clear: the model implies that the first consequence of the anti-inflationary policy now studied will be to increase for a while the rate of inflation. The only way to avoid this conclusion of Equations (159) and (162) would be to assume that r_1 is not larger than r_0, i.e. the unexpected announced policy change would be to *decrease* the rate of interest (for instance, $r_1 = r_0 > r_M$ with $g_0 = g_0^*$ and $g_t = g_M + (r_0 - r_M)\omega^t$ for $t > 0$ would satisfy (159) and (162), resulting in $g_t \leqslant g_0^*$ for all t).

Let us reflect on the nature of the paradox so reached: the mathematical study of our model contradicts the heuristic argument given in Subsection 4 above. It also contradicts the heuristic reaction which the reader might have at this point, namely that simply announcing a progressive decrease in the interest rate cannot really be an anti-inflationary policy move. Since such heuristic argument and reaction make sense, something probably went wrong in the modelling.

Is it the hypothesis that the contemplated policy announcements were credible, and if so, why were they incredible? Is it that the strict hypothesis of rational expectations was too extreme to lead to reliable results for the problem under discussion, because of the consequences for inflation during the initial phase of implementation of the policy? It is disturbing to realize that this second question could be stated; indeed, when rational expectations are assumed in macroeconomics, it is usually within frameworks that are intrinsically more complex than the one discussed here, frameworks in which it would often be difficult to detect, even *ex post*, the consequences showing that something went wrong in the modelling.

Perhaps our exercise simply reveals that the strict hypothesis of model-consistent expectations should not be justified just by the observation that private expectations are actually endogenous in the problem discussed. This remark is worth noting. We may have to remember it in the broader context of Section 4.2 of Chapter 9.

Let us still make the comment that our simple example belongs to the large category of cases in which the long-run features of the solution are the same whether expectations are assumed to be adaptive, hence backward looking, or rational and forward looking. In contrast the short-run features usually differ. Some readers may think that the example chosen here looks like a counter-example to the proposition that the short-run features of the rational expectations solution would necessarily be more persuasive than those of the adaptive expectations solution.

4.4. Instrument rules and target rules

Turning to a different subject, we are now in a position to understand how the approach used in the three preceding sections relates to that discussed in Part 3 of the chapter. This was clearly shown by Lars Svensson[76] who distinguished between instrument rules and target rules. Here we shall draw from his article.

The aggregate demand approach to macroeconomic policies, in Chapter 7 as well as in this part of Chapter 8, aims at finding out the values to be given to some instruments in order to reach some results. We may say that it looks for “instrument rules”. For instance, at the end of Section 4.1 our objective was to find the time profile to be imposed on the nominal interest rate in order to permanently reduce the inflation rate by a given amount η; for a particular specification Equations (125) provided the answer; we may say they defined the rule for the interest rate, taken as the instrument of macroeconomic policy. But we did not discuss the “target rule”, which was to reduce immediately and permanently the inflation rate by so much.

On the contrary in Part 3 we posed an objective of the government, which was to minimize the cost function defined by (83) and (84); we then looked for the target, namely for the inflation rate that would best meet the objective,

[76] L. Svensson, Inflation forecast targeting: implementing and monitoring inflation targets, *European Economic Review* (June 1997).

considering the trade-off between inflation and the level of activity, as well as the reactions of private expectations. We may say that we were then looking for a "target rule" and that we deliberately bypassed the problem of finding out the corresponding instrument rule. We bypassed it by the bold assumption that the inflation rate was directly and perfectly controlled.

Clearly, there is no fundamental contradiction between the two approaches, even though simultaneously introducing their respective intricacies may lead to an unwieldy model. L. Svensson precisely shows how the two approaches may match. We shall now follow him.

With our notation we may write his model of the economy as:

$$q_{t+1} - \hat{q} = \beta_1\left(q_t - \hat{q}\right) - \beta_2(r_t - g_{t-1}) + \beta_3 + \eta_{t+1}, \tag{164}$$

$$g_t = g_{t-1} + \alpha\left(q_t - \hat{q}\right) + \varepsilon_t, \tag{165}$$

where η_{t+1} and ε_t are identically and independently distributed random shocks observed in year $t+1$ but unknown in year t. (Svensson also introduces an exogenous variable x_t in both equations; we delete it for simplicity.) His model is similar to ours, as can be seen, for instance, by comparison with (36) and (37). The addition of the random shocks is welcome. The simplifications come from the fact that neither the pressure of demand d_t nor the inflation expectation g_t^e explicitly appear. But the second and third terms on the right-hand side of (164) correspond to a particular specification of d_t within our more general equation (58); similarly the first term on the right-hand side of (165) corresponds to our Equation (6) for adaptive inflation expectations with $b = 0$. The nominal rate of interest r_t is clearly meant to be the policy instrument.

Since (164)–(165) define a very simple model of the economy, it is important to wonder whether it does not miss something which would be really essential. In particular changes in the interest rate affect output with a lag of just one period and the inflation rate with a lag of just two periods (remember that, with our notation, g_t is the rate of inflation between periods t and $t+1$). These simple lags would be much too short if the period was viewed as being about a quarter long, as we assumed earlier here. But Svensson speaks of years rather than of quarters.

For what follows it is convenient to write the relation between g_{t+1} and r_t, namely:

$$g_{t+1} = a_1 g_{t-1} + a_2\left(q_t - \hat{q}\right) + a_3 - a_4 r_t + \xi_{t+1} \tag{166}$$

with:

$$a_1 = 1 + \alpha\beta_2, \qquad a_2 = \alpha(1 + \beta_1),$$

$$a_3 = \alpha\beta_3, \qquad a_4 = \alpha\beta_2, \tag{167}$$
$$\xi_{t+1} = \alpha\eta_{t+1} + \varepsilon_{t+1} + \varepsilon_t.$$

The new random shock ξ_{t+1} is observed in year $t+2$ but unknown in year t, its expectation being then equal to zero.

Let us now consider the objective function L_t for decisions to be taken in period t. We may rewrite (83) as:

$$L_t = E_t\left[\sum_{\tau=0}^{\infty} \gamma^\tau Z_{t+\tau}\right]. \tag{168}$$

Let us moreover at first assume for simplicity:

$$Z_{t+\tau} = \left(g_{t+\tau} - g^*\right)^2. \tag{169}$$

This is a case in which the policy maker would ultimately care only about the inflation rate g_t and would aim at the target g^*. It is common to say that a "conservative" government would aim at a particularly low value of g^*.

Given the linearity of Equation (166), the decision problem is now "linear-quadratic"; so certainty equivalence holds. Given the form of the quadratic objective function and the fact that the decision (on r_t) in period t will not affect g_t, the optimal policy may be characterized by application of the following recursive decision rule:

$$E_t g_{t+1} = g^*. \tag{170}$$

This rule will indeed also imply equality of $E_t g_{t+\tau}$ to the target g^* for all subsequent $\tau > 1$.

Equation (170) may be said to define the target rule: the inflation rate, expected in period t to occur between periods $t+1$ and $t+2$ must be equal to the target g^*. Svensson accordingly speaks of this rule as requiring "inflation forecast targeting".

The corresponding instrument rule must specify the reaction function for the instrument in terms of current information. Here it is the rule applying to the interest rate r_t given information available in period t. Equations (166) and (170) imply:

$$a_4 r_t = a_1 g_{t-1} + a_2\left(q_t - \hat{q}\right) + a_3 - g^*. \tag{171}$$

A high realized inflation, a high realized degree of capacity utilization or a low inflation target require a high interest rate.

Clearly the same kind of analysis applies if the objective function aims not only at a given (low) inflation rate but also at a satisfactory degree of capacity utilization. In line with (84), we may now replace (169) by:

$$Z_{t+\tau} = \left(q_{t+\tau+1} - \hat{q}\right)^2 + b\left(g_{t+\tau} - g^*\right)^2 \tag{172}$$

b being a positive number.

Given the dependence on r_t, (164) for q_{t+1} and (166) for g_{t+1}, the target rule (170) will then be replaced by:

$$E_t\left(q_{t+1} - \hat{q}\right) + b\left(E_t g_{t+1} - g^*\right) = 0. \tag{173}$$

We may prefer to write it as:

$$E_t g_{t+1} = g^* - \frac{1}{b} E_t\left(q_{t+1} - \hat{q}\right). \tag{174}$$

In other words, the inflation forecast for between year $t+1$ and year $t+2$ should equal the inflation target g^* only if the one-year forecast for the degree of capacity utilization is equal to its optimal value $\hat{q}$ (which is here assumed, according to (165), to be the "non-accelerating-inflation degree of capacity utilization"); if the forecast degree is too low, the monetary authority should accept in year t more inflation between years $t+1$ and $t+2$ than it would have otherwise done. Such a strategy will mean that, instead of adjusting the two-year inflation forecast all the way to the (long-run) inflation target, the monetary authority will accept the prospect of letting it gradually return to that target.

From (164), (166) and (174) we may derive the new instrument rule for r_t. Calculus leads to a result which may conveniently be written as:

$$\beta_2(r_t - g_{t-1}) = \lambda\left(g_{t-1} - g^*\right) + \mu\left(q_t - \nu\hat{q}\right) + \beta_3 \tag{175}$$

with:

$$\begin{aligned} \lambda &= \frac{b}{1+\alpha b}, \qquad \mu = \beta_1 + \alpha\lambda, \\ \nu &= 1 + \left[\beta_1 + \alpha b(1+\beta_1)\right]^{-1}. \end{aligned} \tag{176}$$

The real interest rate $r_t - g_{t-1}$ must be high when either the realized inflation rate or the realized degree of capacity utilization is high. Qualitatively this is the same result as with (171); indeed when b tends to be infinitely large the instrument rule (176) tends to be identical with (171). But the weight μ/λ given to q_t relatively to g_{t-1} is a decreasing function of b; indeed when b tends to zero, λ also tends to zero and the reaction of the nominal interest rate to changes in the inflation rate looks passive, since it just amounts to cancel the impact on the real interest rate.

Readers will also find in Svensson's article and in others of his writings useful complements to our discussions in Part 3 of this chapter.

4.5. About monetary policy rules followed during the 1980s[77]

What are the target rules and instrument rules used by central bankers? The question is natural for macroeconomists. As we shall see later in this chapter, it can be approached thanks to the use of various sets of evidence. In particular we can look at published macroeconomic time series and infer from them by econometric estimation. This is the main mode of investigation used in modern research about what are the policy rules actually followed.

Commenting on this research and on its present outcome, J. Taylor[78] notes that it is common to consider rules using the interest rate as instrument and to look at how the rate responds (i) either to deviations of the money supply from some target, (ii) or to deviations of the exchange rate from some target, (iii) or to weighted deviations of the inflation rate and real output from some targets. The third option, the most frequently investigated, means a rule of the type given by Equation (175). Taylor says that, "in order to capture the spirit" of the outcome of recent research, we may indeed retain this equation for a unit period of a year, with values of 0.5 for λ/β_2 and μ/β_2 and values of 2 per cent per annum for g^* and β_3/β_2: this corresponds to a target inflation rate of 2 per cent; if both the inflation rate and real GDP are on target, the short-term interest rate given by the rule is 4 per cent.

In order to look a little more closely at this recent research and at its relation with the ideas exposed in the preceding section, it is interesting to consider here an article by R. Clarida, J. Gali and M. Gertler[79]. The contribution is in-

[77] For an historical perspective see J. Taylor, An historical analysis of monetary policy rules, NBER working paper No. 6768 (October 1998).

[78] J. Taylor, Discretion versus policy rules in practice, *Carnegie–Rochester Conference on Public Policy* (1993), pp. 195–214.

[79] R. Clarida, J. Gali and M. Gertler, Monetary policy rules in practice: some international evidence, *European Economic Review* (June 1998).

teresting because of its international scope: six countries are studied, Germany, Japan and the US as leading actors of the modern financial world, France, Italy and the UK, as important countries possibly influenced by the leaders. Also, the conclusions are likely to be more robust than those reached in earlier econometric estimations of monetary policy rules, which dealt with periods known to have been subject to changes, even to recurrent hesitations, in the strategies of monetary authorities: the monthly series cover in each case a period during which the rule followed by the central bank could be assumed unchanged; series started in 1979 or later (the terminal date varied from end of 1989 for France, at the date when the European Monetary System became more binding, to end of 1994 for Japan and the US, so avoiding in some cases a recent period which could have been marked by a change of monetary strategy).

The specification and the vocabulary do not match exactly what was seen in the preceding section. A central role is played by what the authors call "a target" r_t^* for the nominal short-term interest rate. But the bank is not supposed to give this value to the instrument r_t. The specification is actually:

$$r_t = (1-\rho)r_t^* + \rho r_{t-1} + v_t, \tag{177}$$

where v_t is an unobservable random shock following a white noise process and the smoothing parameter ρ belongs to the interval [0, 1]. The specification thus admits that banks may deliberately smooth changes in the interest rates. They are often reported to do so and to give good reasons for that[80]: abrupt changes might have destabilizing, even disruptive, effects on capital markets; credibility of the central bank would be jeopardized in case of important policy reversals which might have to follow each other too quickly; credibility also requires consensus on the idea that the policy is right and building such a consensus requires time. Let us note in passing that these reasons for smoothing are objections to the kind of time-separability embodied in the objective function which, in the preceding section, was defined by (168) and (172). More elaborate objective functions are, of course, conceivable in order to deal with the objections, at the cost of leading to less easily interpreted instrument rules.

The authors moreover assume that the target r_t^* is determined according to an equation which we may write here as:

$$r_t^* = \bar{r} + b_p E_t\big(g_{t+n} - g^*\big) + b_q E_t\big(q_t - \hat{q}\big), \tag{178}$$

[80] See a formal discussion of the arguments in A. Caplin and J. Leahy, Monetary policy as a process of search, *American Economic Review* (September 1996).

where n, b_p and b_q are positive parameters, n providing a reference time span on which expected acceleration or deceleration of inflation is supposed to be judged; a 12 months span is actually used in the estimates.

The authors say that plausibility is their main justification for assuming (178). But they also write: “Approximate and in some cases exact forms of this rule are optimal for a central bank that has a quadratic loss function over inflation and output, given the kind of macroeconomic environment we have assumed”, and they quote in particular Svensson (op. cit.). Indeed, the argument developed in the preceding section, starting from the objective (168) with (172) and taking the linearity of (164) and (165) into account, leads to a similar result, to a similar but not quite identical result. Equation (178) now defining the target for the interest rate must be compared with Equation (173), which defined in the preceding section the target for $E_t g_{t+1}$, the inflation rate to be expected between period $t+1$ and period $t+2$ as a result of the policy to be chosen (this was indeed targeting at an inflation forecast).

There are three differences between (178) and (173). The first one, $E_t g_{t+n}$ as against $E_t g_{t+1}$, is immaterial since we saw that n was meant to be something like twelve months and we reported in the preceding section that a meaningful unit period of the model was about a year. The second difference is that, in (178) in contrast to (173), expectations for g_t and q_t are at different horizons; this may be viewed as an oddity of the specification; we may also venture to say that the difference should not have much effect on the conclusions drawn from the econometric results.

The third difference, namely that $r_t^* - \bar{r}$ appears in (178) but nothing of the sort in (173), is more puzzling. In order to reflect on its meaning and implications, we have to know that the econometric model fitted by Clarida, Gali and Gertler is:

$$r_t = \rho r_{t-1} + (1-\rho) b_p g_{t+n} + (1-\rho) b_q q_t + b_0 + \varepsilon_t, \tag{179}$$

where b_0 is a constant, equal to $(1-\rho)\bar{r} - b_p g^* - b_q \hat{q}$, and ε_t incorporates not only v_t but also the effects of the forecast errors on inflation and output. The error term ε_t is assumed serially uncorrelated, in conformity with the same assumption on v_t and with the hypothesis that expectations of central bankers are rational. But the three authors do not exhibit the model on which these expectations are supposed to be formed. The model ought to contain (177) and (178) but also a specification on how g_t and q_t are determined, that is, on how the private economy is supposed to react to variations in the interest rate.

There seems to be some difficulty in deciding how to interpret the results of econometric fits of (179) within the framework proposed by Svensson. The fits do not reveal the objective function of the monetary authorities as a fit on (173) could have done[81]. Neither do they give exactly the instrument rules followed, since g_{t+n} was not known when r_t was chosen. At time t, g_{t+n} was known to depend on r_t and on other observed data, which the fits do not specify but were taken into account. Notwithstanding the ambiguity, we shall follow the three authors in interpreting their results as concerning the instrument rule; they rather speak of the "monetary policy reaction function" but the reaction clearly bore on the instrument, the interest rate[82].

The authors even argue that, on *a priori* ground, their specification (179) of "the reaction function" is superior to (175) and other similarly backward-looking specifications, such as the one proposed by J. Taylor (op. cit.). Their specification does not rely on the idea that central bankers form their expectations adaptively on the basis of a simple unchanging model of the behaviour of the private economy. They write: "central banks consider a broad array of information".

It is now easy to summarize the results obtained by Clarida, Gali and Gertler, accepting here some simplifications. During the period covered, policies followed by the three leaders, Germany, Japan and the US, well fit (179) with an autoregression coefficient ρ equal to about 0.9 (remember that the unit period is the month). The estimated coefficient b_p is larger than 1 (respectively 1.3, 2.0 and 1.8 for the three countries), which means that changes in the expected rate of inflation lead to changes in the same direction not only in the nominal but also in the real interest rate. The estimated coefficient b_q is definitely lower: it is not precisely determined for the US; it amounts to about 0.3 for Germany and to about 0.1 for Japan. The target inflation rate g^* is about 2 per cent for Germany and Japan, 4 per cent for the US. The authors looked whether the central banks of these countries also significantly included in their reactions consideration of other variables, which were introduced as additional regressors in (179): increases in the growth rate of the money stock were found to lead to extra increases in the interest rates; similarly the Bundesbank and the

[81] A test rejecting either the objective function or rational expectations would be provided if a significant effect would be found for some additional regressors which would not appear in the objective function and be known or could be rationally forecast at time t.

[82] Anticipating what will be considered in Chapter 9, we may say that the central banks use a structural model which jointly determines in particular their expectations and their optimal reaction. Equation (179) is meant to be a consequence of this structural model, but it is not the reduced equation which would directly give the reaction as a function of the values taken by exogenous variables of the model. The authors seem to recognize the point when they say that they choose "a weakly restricted specification ... that is sensible in the context of a wide variety of macroeconomic frameworks".

Bank of Japan tend to increase their rates when the Fed do or when their currencies appreciate.

Monetary policies of the three other European countries considered were notoriously influenced by the policy of the Bundesbank and, for France and Italy, by their entry into the hard ERM in 1990. The authors have various ways to take this aspect into consideration. We may simply note here that the fit of (179) is much improved when the German interest rate is added with a coefficient $(1-\rho)b_r$ and that this leads to a lower value for the coefficient of expected national inflation: the couple (b_p, b_r) is (0.5, 0.6), (0.6, 1.1), (0.6, 0.6) respectively for England, France and Italy. Estimation of b_q is not robust except for England where a value of about 0.2 is found.

5 About the Effectiveness of Monetary Policy

The foregoing part was mainly focused on the implications that aggregate demand analysis had for the time patterns to be given to macroeconomic policies. The study proceeded within a fairly simple framework, which cannot be considered as sufficient in this textbook. We shall, however, not move straight to the next step, namely explicit introduction of the labour market. We shall rather make a new pause for the discussion of an issue that has been recurring in various forms during the past fifty years, namely whether and to which extent monetary policy could play an effective role in sustaining the level of activity. Although the main substantial issue in the debates changed, two common questions appeared all through the period: first, whether aggregate demand analysis was useful for guiding monetary policy; second, whether and how observation of aggregate phenomena does reveal the role of monetary policy.

The high point occurred in the 1970s with the conjunction of two theses: first, a substantial thesis, namely that monetary policy was ineffective, in the sense that it could only affect the path of inflation, but not the level of activity, except perhaps for a short while when it had not been anticipated by private agents; second, the methodological thesis that the proper approach to short-term macroeconomics had to assume both that private expectations were rational and that markets cleared, completely and continually: prices had to be assumed fully flexible, which implied that supplies would exactly match corresponding demands. At the time the substantial thesis was perceived by some economists as emerging from the monetarist positions that Milton Friedman had come to take in his long fight against Keynesians; but as we shall see, Friedman's arguments had somewhat evolved; their main thrust was more subtle; moreover, the dogmatism of the methodological thesis was not consonant with Friedman's methodology, which was much more pragmatic and in which reference to facts had always been presented as determinant.

The common wisdom seems now to assert that the substantial thesis of monetary policy ineffectiveness was rejected by econometric tests. This may, indeed, be taken as correct. However, the econometric literature on the effects of monetary policy developed on its own; it deserves all the more attention here as it showed the basic difficulty of deriving persuasive tests from macroeconomic time series alone and as the conclusions now reached by this literature are not easy to summarize.

Such is the range of topics we can and must now cover. We shall do it within seven sections. The first one will briefly survey the monetarist arguments developed during the 1950s and 1960s. Section 5.2 will be devoted to the rational-expectations market-clearing theory of the 1970s; it will focus on a very compact macroeconomic version of this theory. Section 5.3 will consist of a short presentation of the econometric work that directly concerned the thesis asserting that anticipated monetary policy was ineffective. Sections 5.4 and 5.5 will introduce the reader to the econometrics of vector-autoregressions (the VARs). Sections 5.6 and 5.7 will assess what VARs showed about the short-run effects of monetary policy and explain why empirical conclusions about these effects have to be based on more information than just what is contained in aggregate time series.

5.1. The old monetarist position

It may appear somewhat curious that the ineffectiveness thesis was born in the stronghold of monetarism, the University of Chicago, where was teaching M. Friedman who, since the middle 1950s, had been arguing against Keynesian fiscal policies on the ground that money mattered. The fact is less curious if we think in retrospect: first, Keynes himself, and still more systematically quite a few of his early followers, had disparaged monetary policy as unreliable; there were good reasons for questioning such an extreme form of Keynesianism, which by the way was not accepted either by some of Friedman's opponents, like F. Modigliani or J. Tobin; second, in the late 1960s and early 1970s inflation had become a serious concern, so that monetarism naturally attracted interest as providing a simple explanation of a disturbing phenomenon, quite independently of whether monetary policy could help or not for sustaining activity; third, as we saw in Part 3 of this chapter, Friedman was arguing precisely against activist monetary strategies and in favour of a fixed rule for the growth of the money supply. But the old monetarist position is worth remembering in the context of this part before we move to what was called by some economists "Monetarism mark II".

"The roots of monetarism lie in the quantity theory of money, which formed the basis of classical monetary economics from at least the 18th century."[83] Old writers were pointing to the fact that, over the long run, changes in velocity of circulation of money were smaller than those in the money stock. They were moreover arguing that, over the long run, growth in the physical volume of output had to be determined mainly by real factors, so that monetary changes would mainly influence the price level. They were claiming that the observed long-run association between money and prices confirmed that inflation resulted from monetary overexpansion and could have been prevented by proper control of the money supply. (We shall not repeat what was said here in Section 1.3, but simply note that there does not seem to be any fundamental disagreement with the above sentences.)

In his study of the quantity theory of money, Friedman went further with a closer discussion of both the foundations of the theory, in terms of behaviours and market adjustments, and its empirical validation, to which he devoted a lot of attention (mainly in collaboration with A. Schwartz). The philosophy of economic policy that emerged from this work is based on a number of convictions among which we must clearly distinguish those concerning the functioning of the economic system from those regarding the knowledge and behaviour of policy makers. (Some convictions are undoubtedly also due to Friedman's deep adherence to the overall free-market philosophy prevailing from about 1930 on in "the Chicago school".)

In his analysis of the economic system Friedman does not seem to have ever given much credit to the idea of short-run market clearing. He indeed acknowledged the role of market disequilibria in the short-run, for instance, in his discussion of the Phillips curve and in his definition of the natural rate of unemployment, to which we shall refer in the next part. It was only in the long run that the trade-off between inflation and the degree of utilization of capacities would disappear. Indeed, in the long run capacities, including those of labour, would be efficiently used: Friedman always expressed unfailing confidence in the long-run stability of the spontaneous outcome of the market system.

In early discussions about short-run phenomena he was disputing not so much the logical structure of Keynesian models as their practical usefulness,

[83] P. Cagan, Monetarism, in: Eatwell et al. (eds.), *The New Palgrave Dictionary of Economics* (1987), op. cit.

particularly in comparison with another simple model in which the money supply would determine, with some lag, aggregate nominal income (the price level moving in response to various tensions, similar to those of the short-run Phillips phenomenon). He then thought that Keynesians were over-estimating price rigidities by giving inappropriate assessments about the sizes and supposed stability of various coefficients[84]: the Keynesian multiplier was said to be more unstable and unreliable than in particular the velocity of circulation of money. The latter conclusion was derived from a correlation analysis between three variables, the value of households' consumption, the value of autonomous demand (investment *plus* government expenditures) and the money stock, on US annual time series for sixty years. But his empirical analysis also led him to conclude that the adjustment of nominal incomes to an increased rate of monetary growth involved lags that were long and variable[85].

Friedman's positions about monetary policy seem to be mostly due to his intuitive lack of confidence, first, in the ability of economists to disentangle the complexities of the short run, a dimension which he deliberately neglected in his later work, second and probably more, in the practical advisability of asking politicians, or even central bankers, to conduct macroeconomic stabilization policies: the effects of corrective actions would come long after the diagnosis and be uncertain; supposedly stabilizing policies could turn out to be destabilizing; the practice of active stabilization would be likely to deteriorate the long-run performance of the economy. A fixed rate of increase of the money supply would, on the whole, be preferable and remove the uncertainty for investors of the unexpected effects of discretionary policies.

Reading the writings of what was called in the 1970s "the macro rational-expectations school" or "monetarism mark II" we cannot avoid the feeling of an abrupt change in the vision of the economic system and of the formation of monetary policy. Market clearing would occur at any time rather than as a long-run tendency. Monetary policy would indulge in unexpectedly changing the growth rate of the money supply. An element of continuity between the monetarist visions might be searched in the role of expectations, but it is tenuous.

[84] See M. Friedman and D. Meiselman, The relative stability of monetary velocity and the investment multiplier in the United States, 1897–1958, in: *Stabilization Policies* (Prentice-Hall, Englewood Cliffs, 1963); see also more generally his positions in R. Gordon (ed.), *Milton Friedman's Monetary Framework: A Debate with His Critics* (University of Chicago Press, 1974).

[85] A natural question was to ask how long would, under such conditions, the transition be from the short run to the long run efficient state. In his article, The role of monetary policy, *American Economic Review* (March 1968), Friedman ventured to estimate that in the USA it would require "say, a couple of decades" (p. 11).

Indeed, expectations played a part when Friedman, in the late 1960s, explained how the short-run trade-off between inflation and unemployment was transitory, the long-run Phillips curve being “vertical”: in the short run agents do not immediately adjust to the change in the inflationary environment; neither the frequency nor the size of price revisions are fully adequate to what is the new rate of inflation; but progressively agents revise their pricing behaviour, simultaneously with the trend involved in their price expectations; maintaining a high and constant degree of capacity utilization would require a permanent acceleration of inflation, which would eventually be self-defeating. The phenomenon so perceived is very much in the spirit of adaptive expectations, as introduced in Part 1 of the chapter.

But the argument, which had undisputable explanatory power in the context of the late 1960s when inflation was accelerating, was pushed to the extreme by proponents of rational expectations. The claim was indeed that anticipated changes in monetary policy would not alter real variables even in the short run: such an expected change would give agents the incentive to adjust accordingly (i.e. immediately and fully) the levels of all expected and actual nominal variables; this response would neutralize any intended real effect of the monetary change.

So formulated the argument would seem to apply to any framework under rational expectations. Anticipated monetary policy would have no real effect; such effects could only be recorded if some misperception would occur about either what monetary policy exactly is or what its effects ought to be. We shall see in Section 5.3, first, that the assumed consequence of rational expectations is restrictive and, second, how econometricians built tests of the property so asserted and found it did not fare well.

At the time, attention was also given to a more purely theoretical problem: how could misperception and the resulting equilibrium be formalized in models that would rigorously deal with the circularity following from the rational expectations hypothesis strictly interpreted? Such models were worked out and provided the basis for a market-clearing rational-expectations theory of short-run macroeconomic fluctuations. Considering the importance attached to the research in question and to its outcome, we are now turning to it.

5.2. A market-clearing theory

The idea that misperception of movements in relative prices could explain the short-run correlation of changes in output and in the price level was particularly argued in the 70s, following the lead of R. Lucas[86]. It was illustrated on small models with full market clearing and rational expectations; it became an important theme in several macroeconomic textbooks. A simple version of such models will now be considered[87].

1. The economy is divided into a number of sectors or rather "islands", according to the familiar denomination in the literature, because these sectors do not trade with each other in the main good. Let z be the index of islands ($z = 1, 2, \dots, Z$). In each island the representative producer chooses output in period t as a function of its perceived price relatively to the average price thought to hold in the economy (we may say, for instance, that inputs for future periods will have to be bought at the average price). More precisely, let $y_t(z)$ be the logarithm of output in z and $p_t(z)$ the logarithm of the price at which output is sold. Let $E_z\bar{p}_t$ be the expectation held in island z in period t about the log-price that is simultaneously holding on average in the economy. The supply function is written as:

$$y_t(z) = y^N + \alpha\big[p_t(z) - E_z\bar{p}_t\big], \tag{180}$$

where α is a positive coefficient and y^N an exogenously given normal level of log-output.

The crux of the argument lies in the idea that $E_z\bar{p}_t$ is somewhat erroneous, whereas $p_t(z)$ is exact, since it is observed on the market of the island at the time of the supposedly instantaneous production. The expectation $E_z\bar{p}_t$ is, however, rational given the information available to producers in island z. More will be said about this information in a moment.

Supply in island z has to be equal to demand, since $p_t(z)$ is such as to clear the market. The value of demand is exogenously given. We might say that it results in part from fiscal policy; but the literature rather refers to monetary policy and poses:

$$p_t(z) + y_t(z) = m_t + v + \varepsilon_t(z), \tag{181}$$

where m_t is the logarithm of the quantity of money in the economy, whereas $\varepsilon_t(z)$ is a random shock with a nil mathematical expectation. So, the demand

[86] The intuition was first formalized in R. Lucas, Expectations and the neutrality of money, *Journal of Economic Theory* (April 1972).

[87] This draws from R. Lucas, Some international evidence on output-inflation tradeoffs, *American Economic Review* (June 1973). More precisely the model is a simplified version of the one presented in R. King, Monetary information and monetary neutrality, *Journal of Monetary Economics* (March 1981).

Equation (181) follows from a strict quantity theory of money[88], perturbed only by a shock specific to the island and the period. (Note that the constants v and y^N are chosen here for simplicity to have the same value everywhere, so that all islands have equal size.)

Given these supply and demand functions, the log-price in island z has to be:

$$p_t(z) = \frac{1}{1+\alpha}\left[m_t + v + \varepsilon_t(z) - y^N\right] + \frac{\alpha}{1+\alpha} E_z \bar{p}_t. \tag{182}$$

The monetary variable m_t is assumed to be generated by a random walk:

$$m_t = m_{t-1} + \eta_t \tag{183}$$

the shocks η_t having a nil mathematical expectation. The $Z+1$ random elements of the system thus are η_t and the $\varepsilon_t(z)$. They are assumed to be serially independent and normally distributed, with variances σ_η^2 and σ_ε^2; the $\varepsilon_t(z)$ are independent of η_t and such that

$$\sum_z \varepsilon_t(z) = 0. \tag{184}$$

This last equation is assumed for convenience and would only hold as an approximation if the $\varepsilon_t(z)$ were mutually independent and Z was large.

In order to complete the model, we just have to be precise on the formation of expectation $E_z \bar{p}_t$, that is, on the information available to producers of island z in period t. This is made of three parts:

- knowledge of the economic laws ruling the economy: markets clear; expectations are rational; supplies and demands are formed according to (180) and (181), where y^N, α and v are known; the quantity of money follows a random walk; stochastic elements have the properties spelled out above;
- the quantity of money of period $t-1$ is known, but not yet that of period t;
- the log-price $p_t(z)$ is known, but not prices in other islands.

Let us pause and reflect on the representation given by the model to be used here. Whereas we are ready to accept (180) as a convenient approximation to

[88] Except for the shock $\varepsilon_t(z)$, nominal income instantaneous adapt to changes in the money supply. No attention is given to the “long and variable” lags found by M. Friedman in his empirical study of the relationship.

cases in which by hypothesis suppliers are not demand constrained, we cannot but wonder about the realism of the demand function. Taken literally it would imply that demand changes in proportion to the money supply from period to period (quarter to quarter?), except for purely random fluctuations; it would also imply an instantaneous price elasticity equal to -1. Such properties are hard to accept as true. Similarly, assuming that consumers have no knowledge of the money stock except one period later, but immediately behave as if they knew m_t looks strange. But exactly matching the facts is not required; so, let us remember only that demand is assumed to be subject to two kinds of shocks: local shocks $\varepsilon_t(z)$ and a universal shock η_t, neither of which are directly observable by suppliers. Moreover, those suppliers are assumed to have an excellent and confident knowledge of the model.

2. Now we are going to show that there is an equilibrium of the system defined above, an equilibrium in which[89]:

$$p_t(z) = m_{t-1} + v - y^N + \pi\left[\eta_t + \varepsilon_t(z)\right] \tag{185}$$

with a conveniently chosen coefficient π, which will be determined from known elements (precisely from α, σ_η and σ_ε, as we shall see).

Equation (185) shows that, within that equilibrium, observation of $p_t(z)$ entails knowledge of:

$$\xi_t(z) = \eta_t + \varepsilon_t(z). \tag{186}$$

The new random variable is normal with mean zero and variance $\sigma_\eta^2 + \sigma_\varepsilon^2$. Moreover, the theory of the bivariate normal probability distribution tells us that, conditional on the observation of $\xi_t(z)$, the mathematical expectation of η_t is equal to:

$$E(\eta_t/\xi_t) = \frac{\sigma_\eta^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}\xi_t(z). \tag{187}$$

[89] We are here indeed in the favourable linear case, in which it is fairly easy to find the interesting forward-looking solution to a rational-expectations equilibrium problem. Most non-linear cases are much harder to solve. This was realized in particular with the example of the initial 1972 Lucas' article (op. cit.): the solutions of the key equation were the object of further research by Lucas and others who showed that the "neutrality of money" did not hold in general. See P.-A. Chiappori and R. Guesnerie, The Lucas equation, indeterminacy and non-neutrality: an example, in: P. Dasgupta et al. (eds.), *Economic Analysis of Markets and Games*, Essays in Honor of Frank Hahn (MIT Press, 1992).

Equation (185) implies:

$$\bar{p}_t = m_{t-1} + v - y^N + \pi \eta_t \tag{188}$$

and Equation (187) then implies:

$$E_z \bar{p}_t = m_{t-1} + v - y^N + \frac{\pi \sigma_\eta^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}\big[\eta_t + \varepsilon_t(z)\big]. \tag{189}$$

In other words, the rational expectations of suppliers, who like us have derived Equations (185) and (189), lead them to decide on the basis of:

$$E_z \bar{p}_t = \big[\sigma_\varepsilon^2 + \sigma_\eta^2\big]^{-1}\big[\sigma_\eta^2 p_t(z) + \sigma_\varepsilon^2\big(m_{t-1} + v - y^N\big)\big]. \tag{190}$$

Taking account of Equations (188) and (189), we now see that (180) becomes:

$$y_t(z) = y^N + \frac{\alpha\pi\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}\big[\eta_t + \varepsilon_t(z)\big]. \tag{191}$$

Hence, with (185):

$$p_t(z) + y_t(z) = m_{t-1} + v + \left[1 + \frac{\alpha\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}\right]\pi\big[\eta_t + \varepsilon_t(z)\big]. \tag{192}$$

It then emerges that Equation (181) will be fulfilled, and so (185) will lead to an equilibrium, if and only if:

$$\pi = \left[1 + \frac{\alpha\sigma_\varepsilon^2}{\sigma_\varepsilon^2 + \sigma_\eta^2}\right]^{-1}. \tag{193}$$

Equation (191) will then implies:

$$y_t(z) = y^N + (1 - \pi)\big[\eta_t + \varepsilon_t(z)\big]. \tag{194}$$

Let us consider the rational-expectations, dynamic, stochastic equilibrium defined by (185), (193) and (192). Since π is positive and smaller than 1, output and price respond to shocks in the same direction, essentially because suppliers are unable to decompose movements in demand into local and universal components. An increase in σ_ε^2, which reflects the variability in the local factors of demand, increases the quantity response to (unexpected) shocks in the money supply. Observers like econometricians would find a positive correlation between aggregate output and the price level, even though agents are responding only to perceived relative price changes.

Before we move to a test of the ideas incorporated in the model, we may heuristically wonder whether the model just considered offers a persuasive explanation of the covariation of prices and quantities. Can persuasiveness follow from such analytical sophistication applied on a rather bold caricature of economic conditions? Can it do so when more directly observable reactions to market disequilibria are also providing explanation of the covariation?

3. Boschen and Grossman[90] performed a test about the effects of the private information assumed to underlay the behaviour described in the model. In order to do so, they first introduced three changes, so as to make the model suitable for a confrontation with actual data. First, they accepted the idea that, apart from the impact of misperception of relative prices, output evolved around an autoregression; they thus replaced (180) by:

$$y_t(z) = \alpha\big[p_t(z) - E_z\bar{p}_t\big] + a(L)y_{t-1}(z) + c, \tag{195}$$

where $a(L)$ is a polynomial of the lag-operator L and where the constant c plays the same role as y^N in (180).

Second, they explicitly allowed for a systematic monetary policy in which the monetary growth rate responded to the past behaviour of aggregate output. The authors then replaced (183) by:

$$m_t = m_{t-1} + b(L)y_{t-1} + \eta_t. \tag{196}$$

Third, they wanted to be precise on what was known about the quantity of money. They pointed to the fact that preliminary monetary data were issued with a short lag, but were later revised several times over a period of many months. Thus it was irrealistic to assume that, with a unit period covering a quarter, say, the private sector had no knowledge about the contemporaneous money stock and behaved as if it only knew its value three months earlier. The authors then posed the following assumption: although producers do not know exactly m_t during period t, they know its value $\hat{m}_t$ according to the preliminary data. Similarly producers do not know exactly the monetary growth $gm_t = m_t - m_{t-1}$ but its value $g\hat{m}_t$ according to the data available in period t. (The

[90] F. Boschen and H. Grossman, Tests of equilibrium macroeconomics using contemporaneous monetary data, *Journal of Monetary Economics* **10** (1982) 309–333.

authors remarked that monetary authorities themselves did not know the exact value of m_{t-1} when they decided about m_t; but the errors in the data could be seen as just a part in the random disturbance η_t.)

Arguing along lines similar to those followed by us in the preceding section, the authors solved the revised model; they showed that market clearing and rational expectations implied a zero correlation between the output innovation $y_t - a(L)y_{t-1}$ and each one of the contemporaneous or past perceived monetary growth variables $g\hat{m}_{t-\tau}$ for $\tau = 0, 1, 2, \ldots$ Market clearing and rational expectations also implied a positive correlation between output innovation $y_t - a(L)y_{t-1}$ and at least one of the errors $gm_{t-\tau} - g\hat{m}_{t-\tau}$ in the contemporaneous or past announced monetary growth (these errors then corresponding to unperceived actual movements in the supply of money).

Using the quarterly US data for the period 1953 to 1978, the authors concluded that both properties had to be rejected: current and lagged values of the contemporaneous published measure of money growth were significantly correlated with innovations in industrial production, as well as in employment; the joint hypothesis of a zero correlation between all the errors in preliminary monetary data and the same innovations in output and employment could not be rejected. To translate these results into the language most commonly used in the literature here at stakes, we may say that the tests suggest that anticipated money growth causes changes in output and employment, whereas no such effect of unanticipated money growth is revealed, precisely the reverse of what was asserted by the model. The authors conclude: "these tests provide strong evidence against the reality of the equilibrium approach". This conclusion obviously deserves to be confronted with a broader literature.

5.3. Effectiveness of anticipated money growth

At the end of Section 5.1 we met the idea that, under rational expectations, anticipated changes in monetary policy would give to private agents just the required incentive for them to neutralize by their decisions any intended real effect of the policy. When the idea was put forward, it first appeared to be independent of the market-clearing assumption. Research however soon showed that even small elements of price stickiness would invalidate the ineffectiveness proposition. Particularly persuasive in this respect was an article by S. Fischer, of which the basic argument will be given here thanks to a simplified version of the model used[91]. This will provide a background for later turning to the econometric tests of the ineffectiveness proposition.

1. Fischer explores the consequences of the fact that a number of agents contract in nominal terms for periods that are much longer than the time it takes the monetary authority to react to changing economic circumstances. A model is set up in which private agents are assumed to have the same information as the monetary authority: they know in particular the past shocks experienced by the economy up to the latest period and can take them into account in their decisions; they also know the monetary rule, hence the money supply. Fischer then examines two cases of labour contracts; in the first case contracts are negotiated every period (every quarter? Perhaps rather every year, but then monetary policy must not be decided more frequently); in the second case the contracts are negotiated every other period, on odd periods for half of them, on even periods for the other half. The model shows that, accepting all other hypotheses of the aggregate models used in the rational-expectations market-clearing theory, the ineffectiveness proposition does not apply to the second case.

Let the log-linear supply and demand functions be:

$$y_t = y^N + [p_t - w_t + c] + u_t, \tag{197}$$

$$y_t = m_t + v - p_t. \tag{198}$$

They so correspond to (180) and (181) with w_t, (the logarithm of) the wage rate, replacing the expected value of p_t except for the constant c, with the coefficient α being taken as equal to 1 for simplicity, and with the disturbance u_t now concerning the supply and following a first order autoregression:

$$u_t = \rho u_{t-1} + \varepsilon_t, \quad |\rho| < 1 \tag{199}$$

the ε_t being a pure white noise, the supply shock or "innovation". We realize that such specifications of the supply and demand functions are disputable; but accepting them helps in conveying the main point made by Fischer. Market clearing then requires:

$$2p_t = w_t - c + m_t + v - y^N - u_t. \tag{200}$$

[91] S. Fischer, Long-term contracts, rational expectations, and the optimal money supply rule, *Journal of Political Economy* (February 1977). Other elements of stickiness are examined, and lead to a broadly similar conclusion, in another article of the same journal issue: E. Phelps and J. Taylor, Stabilizing powers of monetary policy under rational expectations.

The labour contracts aim at an expected real wage equal to the constant c (the model does not consider the labour market nor a requirement for its clearing). In the first case of contracts negotiated at the beginning of every period, when m_t and u_{t-1} are known but not ε_t, we write:

$$w_t = p_t^e + c. \tag{201}$$

Introducing this in (184) and taking expectations we find:

$$p_t^e = m_t + v - y^N - \rho u_{t-1}. \tag{202}$$

The three last equations then imply:

$$2(p_t - w_t + c) = -\varepsilon_t \tag{203}$$

and (181) leads to:

$$y_t = y^N + \frac{1}{2}\varepsilon_t + \rho u_{t-1}. \tag{204}$$

The level of activity is independent of the monetary policy, which by assumption in the model is exactly anticipated. The ineffectiveness proposition holds. Output fluctuates because, first, of the contemporaneous supply shock and, second, of the impact of past shocks on price expectations and wages. But with the assumed supply and demand equations, this will occur no matter what the monetary authority will decide. Their action will be fully absorbed in price level changes.

The second case needs an argument that is a little more involved. Indeed, the wage contracts no longer lead to exactly (201) but rather to:

$$w_t = \frac{1}{2}\left[p_t^e + {}_{t-1}p_t^e\right] + c, \tag{205}$$

where p_t^e is the price level of period t as expected at the beginning of the period t before ε_t is known, whereas ${}_{t-1}p_t^e$ is the same price level but as expected at the beginning of period $t-1$, at the time when the old two-period contracts were negotiated. (Note an important feature: the two-period contracts are not contingent on what will be observed at the beginning of the second period.) Introducing the wage-setting rule in (200) we now get the following determination of p_t:

$$2p_t = \frac{1}{2}\left[p_t^e + {}_{t-1}p_t^e\right] + m_t + v - y^N - u_t. \tag{206}$$

We also note that writing this equation amounts to maintaining the assumption that the market for goods clears.

On the basis of the equation, expectations are formed, first at the beginning of period $t-1$ when the expectations of p_t and p_t^e are both equal to ${}_{t-1}p_t^e$, second at the beginning of period t when earlier expectations are taken as given. Expectation at the beginning of period $t-1$ leads to:

$$ {}_{t-1}p_t^e = {}_{t-1}m_t^e + v - y^N - \rho^2 u_{t-2}, \tag{207}$$

where ${}_{t-1}m_t^e$ is the expectation formed about the money supply of period t before u_{t-1} and ε_{t-1} are known. Expectation of p_t on the basis of (190) at the beginning of period t leads to:

$$p_t^e = \frac{1}{3}{}_{t-1}p_t^e + \frac{2}{3}\left(m_t + v - y^N - \rho u_{t-1}\right). \tag{208}$$

Taking now into account Equations (197), (199), (200) and (205) to (208) we can solve the model. We find:

$$y_t = y^N + \frac{1}{3}\left[m_t - {}_{t-1}m_t^e\right] + \frac{1}{2}\varepsilon_t + \frac{2\rho}{3}\varepsilon_{t-1} + \rho^2 u_{t-2}. \tag{209}$$

In comparison with what was found in the first case, we see that the policy rule matters, since the error in the expectation of the money supply plays a part in the determination of the level of activity. The point is that, although exactly anticipated at the beginning of period t, the outcome of the policy rule cannot be so well anticipated earlier when longer term wage contracts are negotiated. This creates a situation in which the monetary authority can influence not only prices but also activity, even though the monetary rule may be exactly known by private agents. The ineffectiveness proposition no longer applies.

A policy rule will naturally specify how to take account of the latest revealed supply shock. So we may formalize the rule as defined by the choice of a number a, which will characterize the policy according to the formula:

$$m_t = m_{t-1} + a\varepsilon_{t-1}. \tag{210}$$

This rule will lead to the expectation m_{t-1} for m_t at the beginning of period $t-1$, hence to:

$$y_t = y^N + \frac{1}{2}\varepsilon_t + \frac{1}{3}(a + 2\rho)\varepsilon_{t-1} + \rho^2 u_{t-2}. \tag{211}$$

So, output stabilization will depend on the rule. The most effective stabilization in this respect will be reached with $a = -2\rho$. It will result in:

$$y_t = y^N + \frac{1}{2}\varepsilon_t + \rho^2 u_{t-2}. \tag{212}$$

We note that, in comparison with the first case in which (204) applied, output stabilization is better achieved because of the impact of the shock ε_{t-1} on the revision of price expectations at the beginning of period t and because of the clever choice of the policy rule.

2. Reflecting on this model of S. Fischer and on the model of the previous section should permit us to better understand: (i) how the logic of the rational-expectations market-clearing approach operates, (ii) what the ineffectiveness proposition exactly means, (iii) how it could be deductively proved to hold or be invalidated, once we would be sure about the realism of a particular model and of the rational expectations hypothesis within it.

But the question remains to know whether available macroeconomic models of that type are just made as providing exercises for students or whether they capture anything that would be significant in reality, and if so what it is. The fact that this tantalizing question is seen by many economists as remaining unanswered, more than any other reason, may explain why the approach attracted definitely less attention from theorists in the 1990s than it had during the two preceding decades, and why most applied works in short-run macroeconomics still rely on assumptions implying that somehow markets are not "in equilibrium".

Attitudes with respect to the question also depend on the reading made of empirical evidence, the macroeconomic evidence reported at the end of Section 5.2, the one to be now reported, but also all the microeconomic evidence which otherwise exists about actual short-term operations in market economies, where, for instance, such rigidities as those implied by long-term non-contingent contracts go definitely farther than is recognized even in Fischer's model.

Our reflection should also bear on the strict interpretation that we give of the rational expectations hypothesis when we say that expectations have to be congruent with the model in which they appear. After the two models just discussed, but also after the demand-driven model introduced in the last part of Section 4.3, we see that it is not innocuous to assume that private agents know the model and believe it is the true model. It is not innocuous because it leads to results that may greatly differ from those reached with, for instance, adaptive expectations; it leads to results that may moreover be difficult to explain heuristically.

Common sense suggests that a softer interpretation of the rationality of expectations would lead us closer to actual facts: in normal times agents would adaptively revise their expectations, in conformity with their rather simple-minded representations of macroeconomic phenomena; in exceptional times, when expectations have to be wholly reconsidered, the rational expectations hypothesis may be the only available tool for thoughts, but then within a model that has to properly deal with the exceptional circumstances. There is little chance for such a model to exist in advance before the circumstances are exactly revealed. With such a softer interpretation of the rationality of private agents, the strict rational expectations hypothesis could very well lead to misleading assessments about the effectiveness of policy rules.

3. R. Barro defined the structure of an econometric test of the hypothesis that anticipated monetary policy would have no effect on output[92]. This structure was accepted, with only few changes, in a number of subsequent tests. The basic idea is to characterize the anticipated quantity of money, or the anticipated rate of growth of this quantity, as a latent unobserved variable that can be inferred from a number of observable determinants, the inference being based on a past regression of the quantity of money on the determinants in question.

In order to write the equations used by the econometrician, let us define for definiteness y_t as real output in period t, y_t^N as the "natural" level of output or more exactly its trend, x_t as the rate of growth of the money supply, z_t as a vector of exogenous variables. The econometrician will typically proceed in two steps. He will first estimate the vector n of coefficients in a regression:

$$x_t = nz_t + \varepsilon_t. \tag{213}$$

Let n^* be the estimate. The econometrician will then define:

$$x_t^e = n^* z_t \tag{214}$$

and use this variable as a good proxy for the anticipated rate of money growth. He will finally fit an equation such as the following:

$$y_t = y_t^N + \sum_{\tau=0}^{h} a_\tau x_{t-\tau}^e + \sum_{\tau=0}^{h} b_\tau \left(x_{t-\tau} - x_{t-\tau}^e\right) + u_t. \tag{215}$$

[92] See, for instance, R. Barro, Unanticipated money growth and unemployment in the United States, *American Economic Review* (March 1977).

The unknown coefficients a_τ and b_τ, meant to concern the effects of respectively the anticipated and unanticipated money growth, will be the object of estimations and/or tests. More precisely the ineffectiveness of the anticipated monetary policy will mean $a_0 = a_1 = \cdots = a_h = 0$ and be tested as such.

There are a number of questions on how to perform such a test. Perhaps the most relevant one here is to know which variables should be taken as determining changes in the money supply according to (213), that is, which variables private agents know as being considered as relevant by policy makers. In one such US study[93] the vector z_t had eleven components, a constant, six lagged values of x_t, three lagged values of the logarithm of the unemployment rate and a contemporaneous measure of the deviation of federal government expenditure from its normal value.

But there are other questions: how long should the horizon h be in Equation (215)? Considering what has already been learned about the lags with which monetary policy could affect the economy we may think of something like 12 quarters. How should the trend y_t^N be specified: as a pure exponential, with possibly breaks in the rate of potential growth, with possibly addition of a stationary random disturbance or even of a random walk? What should be assumed about a possible serial correlation of the disturbances u_t in (215)? Should there be any implication in the econometric procedure of the discrepancy between the proxy x_t^e and the actually anticipated monetary growth? Discussing these various questions might retain our attention for a long time.

Suffice it to say here that, after the initial work of R. Barro, which appeared to give support to the ineffectiveness proposition, further econometric works rather rejected it[94]. This went up to the point of stating that anticipated money growth was just as effective for stimulating output as unanticipated money growth. In other words, the hypothesis that, in (215), $a_\tau = b_\tau$ for all lags τ could not be rejected; moreover the conclusion appeared to be robust to changes in the specification of rational expectations and of the natural trend of output[95].

If universally accepted this conclusion would, of course, strongly limit the importance of expectations about the money supply in the analysis of the real effects of monetary policy. The conclusion would also simplify the vision of

[93] R. Barro and M. Rush, Unanticipated money and economic activity, in: S. Fischer (ed.), *Rational Expectations and Economic Policy* (University of Chicago Press, 1980).

[94] See, for instance, F. Mishkin, Does anticipated monetary policy matter? An econometric investigation, *Journal of Political Economy* (February 1982).

[95] R. Frydman and P. Rappoport, Is the distinction between anticipated and unanticipated money growth relevant in explaining aggregate output, *American Economic Review* (September 1987).

macroeconomists about those effects, since Equation (215) would boil down to:

$$y_t = y_t^N + \sum_{\tau=0}^{h} a_\tau x_{t-\tau} + u_t. \tag{216}$$

It would simplify the macroeconomic vision in one respect, but not in all respects unfortunately. Indeed, the last equation is precisely in the heart of a more permanent problem in empirical assessments about the effectiveness of monetary policy, a problem that was raised about forty years ago and is not yet settled, a problem to which we now turn.

5.4. The empirical approach: preliminaries

1. Published in 1963 *the initial work of M. Friedman and A. Schwartz*[96] was the main source from which Friedman drew arguments in favour of his conviction that changes in the supply of money were the principal cause of changes in money income. This work was empirical; arguing from it Friedman had to be clear on the nature and force of the empirical evidence.

In the tradition of the time at the National Bureau of Economic Research which sponsored this work, great attention was being devoted to the characterization of leads and lags between economic variables, particularly at cyclical turning points. It was found that the rate of change of the money supply exhibited a long lead over nominal income; even the money supply itself was usually leading income by some four months to two years. In the early period of the work such leads may have been taken as decisive evidence of the influence of money on income. But Friedman soon realized that a timing pattern between variables subject to common cyclical fluctuations was weak as empirical proof of a causal link. An economist might build a case showing that the lead could exist even if the quantity of money would just passively react to changes in income. The question was then to know whether there was evidence of an independent influence of the money supply. History of the monetary process was scrutinized; occasions were found in which a purely passive response was most likely; but equally obvious seemed to be instances where the money supply changed for reasons quite independent of past or contemporaneous movements in money incomes: the boom in gold production between

[96] M. Friedman and A. Schwartz, *A Monetary History of the United States, 1867–1960* (Princeton University Press, 1963).

1896 and 1913, the financing of wars, the waves of banking failures or the changes in the regulation of the banking sector.

In presenting the conclusions of his work with A. Schwartz, Friedman was therefore careful and several times explained that the timing evidence was "suggestive but by no means decisive". The following extract from *A Monetary History* shows the type of formulation he came to prefer: "Changes in the money stock... are a consequence as well as an independent source of changes in money income and prices, though, once they occur, they produce in their turn still further effects on income and prices. Mutual interaction, but with money rather the senior partner in longer-run movements and in major cyclical movements, and more nearly an equal partner with money income and prices in shorter-run and milder movements – this is the generalization suggested by our evidence"[97].

We note that, in order to draw general conclusions from his observations, Friedman uses no explicit inferential procedure. Indeed, the crucial word is that the generalization is "suggested" by a whole set of facts. (However, in disputes with other economists, Friedman clearly meant the suggestion to be taken as having great force.) We, as macroeconomists, should not dismiss offhand conclusions so reached; we are too aware of the paucity and limitations of conclusions otherwise obtained in our non-experimental discipline. But followers naturally searched for more formal econometric validations of the conclusions. Keeping the inspiration of Friedman's selection of arguments, they tended to focus not on the main movements exhibited by statistical time series but rather on deviations around these movements, deviations to be explained by independent exogenous shocks that generated them. Let us now look at the econometric work in question.

2. We begin with the case of just two time series on two variables x_t and y_t. For instance, they may the quantity of money (x_t) and nominal income (y_t), or respectively deviations of these variables around their longer-run movements. We shall then simply speak of "detrended money" and "detrended income", but the trend is often meant to represent not only the long-run growth but also fluctuations whose periodicity exceeds a few years. Clearly the method used for detrending may matter; but we shall neglect the difficulty at least for the time being.

Faced with the problem of how to *test whether the variable x_t may be taken as a cause of the variable y_t*, statisticians proposed a method meant to be

[97] This is quoted in the following text, which has an independent interest for the question at issue here: Comment on Tobin, *Quarterly Journal of Economics* **84** (1970) 318–329, following J. Tobin, Money and income: post hoc ergo propter hoc?, the same issue, pp. 301–317.

appropriate for anyone ready to accept that anteriority of erratic movements is a sufficient presumption of causality. In order to stress this condition, they speak of "Granger causality"[98]. Notice that Granger causality might very well be observed if the real cause was a third variable z_t affecting both x_t and y_t, with shorter lags for x_t than for y_t. Notice also that there is no contradiction in accepting the idea that causality, although fundamentally an asymmetric relation, might run both ways: from x_t to y_t and from y_t to x_t.

The standard test of Granger causality starts from two fits of linear equations, the first of y_t on past, present and future values of x_t, the second of x_t on past, present and future values of y_t:

$$y_t = \sum_{\tau=-h}^{h} a_\tau x_{t-\tau}, \tag{217}$$

$$x_t = \sum_{\tau=-h}^{h} b_\tau y_{t-\tau}, \tag{218}$$

where h is a number chosen for the application of the test and so depends on circumstances, the condition being that there are good reasons to think that longer lags and leads would not result in a different assessment. The coefficients a_τ and b_τ are estimated by ordinary regressions on (217) and (218). Causality from x_t to y_t is understood to mean that all the coefficients a_τ in (217) for negative τ would be nil, if it were not for a random noise around the equation: so the true representation of the link between the y_t series (assumed to be the realization of a stationary process) and the x_t series (similarly assumed stationary) may be taken as involving only contemporaneous and past values $x_{t-\tau}$ together with purely random disturbances. In other words, a standard F-test on the joint hypothesis $a_{-h} = a_{-h+1} = \cdots = a_{-1} = 0$ will show whether Granger causality from x_t to y_t has to be rejected or not. A similar F-test for causality from y_t to x_t may be provided by a fit on (218).

Applying this statistical method and considering the quarterly series of the United States for the period 1948 to 1968, C. Sims found a unidirectional Granger causality from the money stock to the gross national product in current value[99] (more precisely, the data could not reject causality from money

[98] For a careful presentation of the background giving meaning to the concept see C. Granger's entry on "causal inference" in Eatwell et al., *The New Palgrave Dictionary of Economics* (1987), op. cit.

[99] C. Sims, Money, income and causality, *American Economic Review* (September 1972). The results were updated and their interpretation in relation to the VAR approach was discussed in C. Sims, Models and their uses, *American Journal of Agricultural Economics* (1989).

to income, but did reject causality from income to money). The result gave a validation of the monetarist thesis.

The main objection against such an interpretation is that it isolates two variables as if their joint evolution was not affected by any other outside influence than unobservable erratic shocks striking at times one of the variables, at times the other. This is a strongly restrictive assumption, which amounts to ignoring a long tradition in the analysis of business fluctuations, a tradition insisting on the need to simultaneously consider the evolution of many variables.

Only in Chapter 9 shall we closely examine the tradition in question, how it led up to macroeconometric models, how it was then attacked by economists inclined to rely on more deliberately empirical methods, how C. Sims in particular explained that it was possible to so deal with several variables and developed for the purpose a technique starting from vector autoregressions (the so-called VAR technique). But we must somewhat anticipate now on what will be studied in the next chapter because the use of simple VARs has a large place in modern discussions about the effectiveness of monetary policy (we shall also see in Part 7 of the chapter how VARs may intervene for the identification of supply and demand shocks).

5.5. Introduction to the VARs

1. Let us now assume that we consider series of not just two but several variables. Let y_{jt} be the series for a variable ($j = 1, 2, \ldots, k$) and y_t be the vector of the observations for the k variables simultaneously. *A vector auto-regression is defined* as a fit on a model explaining the current value of the vector y_t by its past values and by an unobservable vector of disturbances.

There are various ways of specifying the model and hence of defining the vector autoregression. For convenience in this presentation[100] we shall start from the hypothesis that the vector y_t is the realization of a multidimensional stationary process such that:

$$\sum_{\tau=0}^{h} A_\tau y_{t-\tau} = A(L)y_t = \varepsilon_t, \tag{219}$$

where the k-dimensional disturbance vector ε_t is uncorrelated with past $y_{t-\tau}$ (equivalently with past $\varepsilon_{t-\tau}$) and has an identity covariance matrix, where

[100] This is the specification chosen in an article to which we shall particularly refer here: E. Leeper, C. Sims and T. Zha, What does monetary policy do?, *Brookings Papers on Economic Activity*, No. 2 (1996).

moreover A_0 is invertible. Assuming for the time being that we know the $h+1$ square matrices A_τ, we can recursively solve (219) for computing the y_t from the $\varepsilon_{t-\tau}$. The general formula is:

$$y_t = \sum_{\tau=0}^{\infty} C_\tau \varepsilon_{t-\tau}. \tag{220}$$

We assume here, without discussion, that the infinite sum converges, which requires that all characteristic roots of A_0 have modulus larger than 1. Convergence means that the matrix C_τ become infinitely small with increasing τ. In mathematical statistics an equation such as (220) is said to give a moving-average representation of the autoregressive stochastic process ruling y_t.

Great use is made in VAR analyses of the representation (220). Indeed, the element $c_{ji\tau}$ of the matrix C_τ may be taken as measuring the intensity of the effect of the shock denoted ε_i on the variable y_j after a lag of τ periods. The graph of the $c_{ji\tau}$ for successive values of τ is known as the "impulse-response function" of y_j to ε_i: if the graph is everywhere close to the horizontal axis, there is little response; if it is high for small values of τ but quickly decreases toward zero, the response is significant but transitory, and so on. Clearly, for such assessments to the meaningful, we need to know what the i-th component of the disturbance vector ε_t represents. There is an identification problem.

In order to grasp the nature of the problem we must also realize that the representation of the y_t process by Equation (219) is not unique. Let W be any orthonormal matrix of order k, that is, any square matrix such that $W'W$ is equal to the k-dimensional identity matrix I. Replacing $A(L)$ by $WA(L)$ and ε_t by $\eta_t = W\varepsilon_t$ gives another representation of the y_t process and it is equivalent to (219). Note indeed that, the covariance matrix of ε_t being equal to I, the covariance matrix of η_t is equal to $W'W$, hence also to I. The indeterminacy of the representation (219) naturally implies an indeterminacy of the relevant ε_{it} and of the impulse responses: in (220) ε_t may be replaced by η_t, the matrices C_τ being replaced by $C_\tau W'$.

2. C. Sims and his followers have developed *rules for the solution of the identification problem.* They now speak of "identified vector autoregressions" to characterize the class of models they are using. They introduce three sorts of identifying restrictions in order to pin down the definition of $A(L)$ given the behaviour of y_t. First, they use exact linear restrictions on the elements of A_0, usually simply setting some elements to zero. To rely entirely on such restrictions for solving the identification problem, they would need at least $k(k-1)/2$ of them because such is the number of free elements of an orthonormal ma-

trix of order k (the requirement that the symmetrical matrix $W'W$ be equal to the identity matrix imposes $k(k+1)/2$ restrictions on the k^2 elements of W). Actually this number of linear restrictions on A_0 is exactly reached when this matrix can be given a triangular form or equivalently when the autoregression model is truly recursive. This means that the components of y_t can be ordered from 1 to k in such a way that no contemporaneous variable y_{it} with a higher rank i than j appears in the j-th equation of (219), while the various components of the disturbance vector are also uncorrelated with each other. When these and only these restrictions are introduced, we may speak of "recursive identification".

A second sort of identifying restrictions may be found in a prior economic analysis of the meaning of each equation in the system (219), an analysis which restricts the forms that matrices A_τ are likely to have. This sort of restrictions will play a major role in the "structural approach" to be studied in Chapter 9. We shall then explain the introduction of "structural VARs". Here let us simply report that Sims and his followers use probabilistic assertions about elements of $A(L)$, saying that some values (or relations among values) of elements of $A(L)$ are more likely than others. Third, they may also decide to use informal restrictions on the reasonableness of the impulse responses, the elements of $C(L)$ in (220).

Once indeterminacy of $A(L)$ has been removed, it is still necessary to give labels to the components of ε_t; actually, the labelling may emerge as rather natural along with the choice of the identifying restrictions.

In order to briefly focus on the last comment and to better understand how the method operates in practice, let us consider the case of a recursively identified two-dimensional VAR. Let elements of the matrix A_0 be denoted for simplicity by a_{ij} (with $i, j = 1, 2$). Estimation of $A(L)$ will start from a least square two-dimensional regression of y_t on its past values $y_{t-\tau}$ for $\tau = 1, 2, \ldots, h$. Let us write the result of this regression as:

$$y_t = \sum_{\tau=1}^{h} B_\tau^* y_{t-\tau} + u_t^*. \tag{221}$$

B_τ^* being a matrix of regression coefficients and u_t^* the vector of the two residuals in period t.

Four equations will give the values of the four elements of A_0, the estimated values of the A_τ matrices for $\tau > 0$ then being simply $-A_0 B_\tau^*$ and the vector of residuals ε_t^* corresponding to (219) being $\varepsilon_t^* = A_0 u_t^*$. The first equation will follow from the assumed triangular form of A_0, namely: $a_{12} = 0$. The three other equations will involve the observed covariance matrix of the residual vectors u_t^* on the full sample; let us call it Σ: since the true covariance matrix of ε_t is the identity, we naturally require the observed covariance matrix of the residuals ε_t^* to also be the identity matrix; this means:

$$A_0 \Sigma A_0' = I. \tag{222}$$

This gives us the three remaining needed equations on A_0, namely:

$$\begin{aligned} &\sigma_{11} a_{11}^2 = 1, \qquad [\sigma_{11} a_{21} + \sigma_{12} a_{22}] a_{11} = 0, \\ &\sigma_{11} a_{21}^2 + 2\sigma_{12} a_{21} a_{22} + \sigma_{22} a_{22}^2 = 1. \end{aligned} \tag{223}$$

In order to interpret the results following from the solution of these equations it is convenient to consider an auxiliary regression, that of the residual u_{2t}^* of the second equation in (221) on the residual u_{1t}^* of the first equation:

$$u_{2t}^* = \beta^* u_{1t}^* + v_t^*, \tag{224}$$

where β^* is the regression coefficient and v_t^* the residual in period t. Let σ_{vv} be the sample variance of this residual. Knowing that $\beta^* = \sigma_{12}/\sigma_{11}$ we find that the solution of (223) is:

$$a_{11} = [\sigma_{11}]^{-1/2}, \qquad a_{21} = -\beta^* a_{22}, \qquad a_{22} = [\sigma_{vv}]^{-1/2}. \tag{225}$$

It then follows that:

$$\varepsilon_{1t}^* = [\sigma_{11}]^{-1/2} u_{1t}^*, \qquad \varepsilon_{2t}^* = [\sigma_{vv}]^{-1/2} v_t^*. \tag{226}$$

We now see the effect of the order chosen in the recursive identification: except for normalization the two observed "shocks" (also called "innovations") ε_{1t}^* and ε_{2t}^* are respectively equal to u_{1t}^*, residual in the first of the two regressions (221), and v_t^*, residual in the auxiliary regression. The number v_t^* is sometimes called "orthogonalized residual" because it is the remaining part of u_{2t}^* after elimination of the t-component of an orthogonal projection of the vector u_2^* on the vector u_1^*.

To better understand the meaning of this asymmetry, let us think of an example in which y_{1t} would be the stock of money and y_{2t} an indicator of business activity. Then, according to present practice, the first equation in (221) will be read as showing how the monetary authority reacts to changes in business activity; the residual ε_{1t}^* will provide an estimate of the so-called "monetary policy shock". This shock is indeed a deviation from the systematic rule that the monetary authority is found to follow on average; the shock in question is moreover directly identified from the first regression in (221) because of the hypothesis made in the triangularization.

The economist presenting results about this example may say that the monetary authority is assumed not to react to changes in business activity before a lag of one period; according to the argument, the innovation in the monetary series could not in any way be a "consequence" of changes in business activity (see the quotation from Friedman at the beginning of this section). Instantaneous causality could only occur from money to activity and, since no assumption is made concerning the reaction of the business sector, such a causality could be found.

The lack of symmetry with such a recursive identification may not be serious, particularly if the unit period is short, because then the identifying restrictions may not lead to detect any significant contemporaneous reaction. This will occur precisely if a_{21} is found to be close to zero, which means that there is little correlation between the residuals u_{1t}^* and the residuals u_{2t}^*. But if a_{21} is not negligible, we may be seriously concerned by the meaning of the assumed recursiveness. Indeed, in the example just considered, few economists would consider as an established fact that reactions of the business sector to changes in the money supply would be faster than reactions of the monetary authority to changes in indicators of business activity.

More generally a good deal of judgement is clearly required from economists who apply the technique of identified vector autoregressions, and from economists who want to interpret the results so obtained. We shall not elaborate more on this point in the abstract. We shall rather consider whether and how in practice the approach may enlighten queries about the effectiveness of monetary policy.

5.6. The empirical approach: some results

1. A natural starting point concerns evidence on *the relations between money, output and prices*. Does the VAR approach validate the monetarist idea according to which an increase in the money supply has a temporary favourable impact on output, but that its longer-run effect is a proportional increase in the price level? Such was indeed a question already motivating the first application by C. Sims and his initial plea in favour on the VAR approach[101]. Such is also

[101] C. Sims, Macroeconomics and reality, *Econometrica* (January 1980).

the issue underlying the first set of results recently presented in Leeper, Sims and Zha (op. cit.). These authors analyse the monthly US series for the period 1959 to 1996, with M_1 being the measure of the money stock. In fits of models such as (219) the maximum lag h is six months.

Recursive identification is used for the VAR on the three series concerning the quantity of money, the volume of output and the price level, with the order M, y, p, which implies the two following assumptions: the money stock does not react before at least a month to changes in output and prices; output does not react before a month to changes in prices. The residuals of the three equations are respectively interpreted as "monetary policy shocks", "output shocks" and "price shocks".

The graphs of the impulse responses following a monetary policy shock are reproduced here in Figure 6. After a policy shock equal to about 0.4 percent of the money stock, responses during the following 48 months are plotted in each of the three graphs. Those are the responses through the working of the economic system, including the reaction of the monetary authority to past changes in output and prices. The money stock quickly increases in the first year, up to about 1.0 per cent more than its initial level, after what it would remain constant, would it not be for new policy shocks. The response of output is not negligible after a year, but small (less than 0.2 per cent of the initial level of output) and transitory (the impact has practically disappeared after four years). The response of the price level is on the contrary permanent and more important (a 0.4 per cent increase) but slower (less than half of it is recorded within a year). We note that, on the basis of these results, the policy shock would have "non-neutral" effects within four year, since the ration M/P would have increased.

Similar impulse responses to output shocks and price shocks are graphed in the article of Leeper et al. (but not reproduced here). They suggest that an output shock would have a permanent effect on output but no other effect, not even an amplification of the initial impact. A price shock would on the contrary be amplified and depress real output, even though py would increase after a positive shock.

The authors claim that their results are robust, but that after substitution of M_2 for M_1 output has a more persistent response. They present these results as follows: "No method of data filtering changes the fact that monetary aggregates contain substantial variation that past output data do not help to predict, or that this variation in money does help to predict future output. The response of the price level to a money stock innovation is smooth and slow; the response of output is quicker and less sustained".

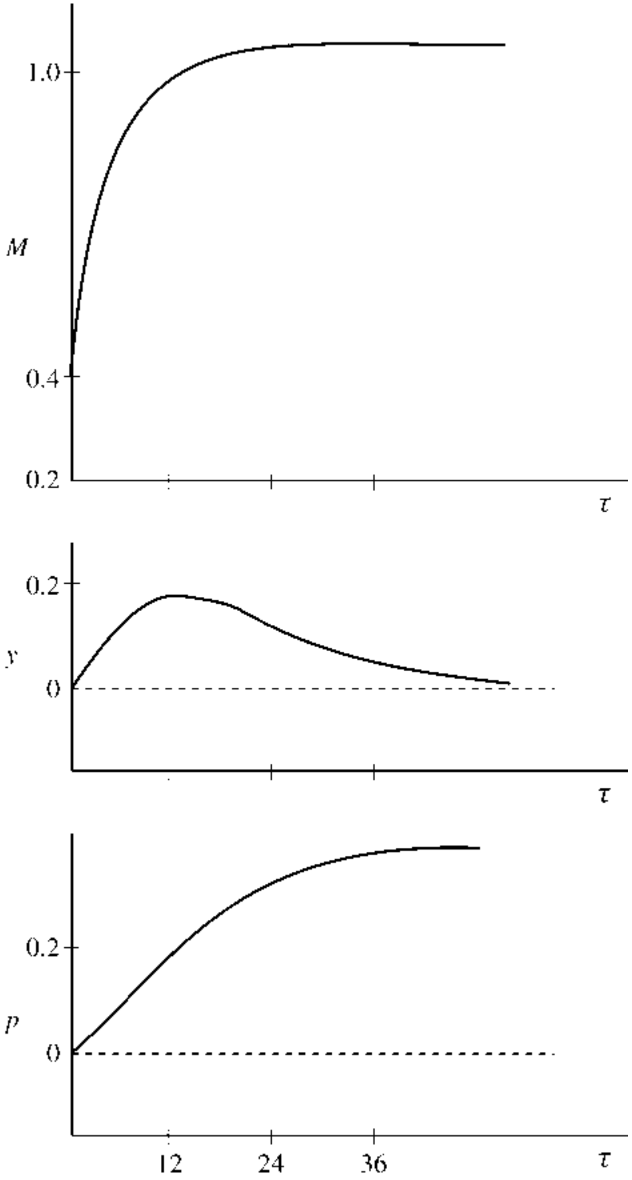

Fig. 6. Impulse-response functions to money shocks.

This is worth remembering. However, the interpretation in terms of policy may be subject to some reservation. Are not the results biased by the fact that shocks in the private demand for money may exist, may then appear as innovations in the time series analysis of the money stock and may have implications of their own on the private demand for goods, hence on output and prices? That such phenomena could be observed quite independently of any change in policy is all the more likely as the monetary authority does not directly control the quantity of money but other instruments: in the US the interest rate on "federal funds" and/or the "non-borrowed bank reserves at the Fed". Indeed, no modern VAR analysis now stops at considering just the three variables M, y and p.

2. *Responses to shocks in the interest rates*, certainly deserve attention. Already in his 1980 article (op. cit.) C. Sims noted that introduction of an interest rate in his VAR analysis was considerably reducing the impulse responses to monetary innovations. Leeper et al. (op. cit.) take the question again in the second set of results they present, based on a four-variable VAR involving not only M, y and p but also the federal funds rate r.

They first use recursive identification with the ordering M, r, y, p. But they argue that two of the impulse responses so found are not reasonable: the response of the interest rate to a positive monetary innovation shows just a very small contemporaneous decrease, immediately followed by a more important increase; the response of the price level to a positive interest innovation is an increase, whereas a decrease would be reasonable. They remark, however, that looking simultaneously at the responses of output (or of prices) to money and interest innovations leads to the explanation that recursive identification is not appropriate in this case and that another set of identifying constraints will lead to reasonable results. So, they turn to the third sort of identifying restrictions they signaled in their methodological preliminaries.

The exact argument leading them to the choice of a new set of identifying restrictions is somewhat involved and can be read in their article. Suffice it to say here that they identify in particular a policy-making sector, from which originate policy shocks; those shocks combine monetary expansion with a corresponding decrease in the interest rate, or monetary contraction with a corresponding increase in the interest rate. Leeper et al. maintain the restriction that, in its choice of M_1, the policy-making sector does not react within the month to private behaviour; but it partly accommodates the impact of this behaviour on the interest rate. The output response to a shock on M_1 is then found to be persistent but small (about 0.1 per cent); the price response has a similar shape to that in Figure 6 but is somewhat weaker (0.3 instead of 0.4 per cent after four years). After the shock the quantity of money evolves about as in

Figure 6, whereas the initial decrease in the interest rate vanishes after about a year. All in all, policy seems to be a little less effective than appeared from the three-variable VAR. Private sector shocks, assumed to contribute to innovations in output, prices and interest rates but not to money innovations now have three dimensions, but those implying price rises systematically elicit a contractionary response from the (nominal) interest rate; most of the observed variation in the interest rate is accounted for by these endogenous responses.

Other econometricians using the VAR approach to the study of monetary policy are not convinced by Leeper et al.'s arguments to the effect that their identifying restrictions would be more "plausible" than alternative ones. The focus of the discussion is the assumption that the money stock does not react before at least one month to changes in private sector behaviour. For instance, in the discussion of the article in question B. Bernanke argues that the choice of the indicator of policy should be based on a scrutiny of the operating procedures followed for implementing decisions of the central bank and that these procedures designate the interest rate as being the main instrument. He then writes "since changes in demand for money or reserves are almost entirely accommodated by the central bank, innovations in money or reserves typically mostly reflect demand shocks rather than policy shocks".

Such being the situation, we find it interesting to look at the results of another VAR study published about at the same time[102]. Christiano et al. analyse quarterly US data of the period 1960–1992, introducing four lags in the vector autoregressions. These authors use alternatively one or the other of two variables as being the policy instrument, either the federal funds rate or the amount of non-borrowed bank reserves at the Fed. They use recursive identification, but with an assumed ordering that radically differs from the ones assumed in Leeper et al. The policy instruments, for instance the federal funds rate, does not come first but rather after three private sector variables: output y, the GDP deflator p and an index of sensitive commodity prices. Their basic assumption is that any contemporaneous correlation between these variables and the policy instrument reflects causation from production and prices to policy, not the other way around.

They find that M_1, which comes still later in the ordering, significantly declines after a shock increase in the federal funds rate; output y responds to the same shock by a sustained and significant decrease after two quarters; the GDP deflator response is roughly flat for a year, after which it slowly declines. Close scrutiny between the impulse response curves and those of Leeper et al. shows

[102] L. Christiano, M. Eichenbaum and C. Evans, The effects of monetary policy shocks: evidence from the flow of funds, *Review of Economics and Statistics* (February 1996).

a number of differences. But the qualitative conclusions are no so different and appear to be "reasonable" in both cases.

At this point we shall not look further into the results of Leeper et al. or Christiano et al. concerning the US economy. The impulse responses estimated for larger sets of variables, including such ones as retail sales, corporate profits or inventories, are interesting, but examination of the larger body of results will better fit in Chapter 9, which will be precisely devoted to the simultaneous study of all main dimensions of business fluctuations. On the other hand, it is now the proper time to evaluate the contribution to our knowledge that may come from the form taken by the empirical approach after the penetration of the VAR technique.

We shall, however, still remark that the features reported above are not special to the US economy. They were roughly confirmed by an investigation involving, besides the US, four other OECD countries (France, Germany, Japan, the UK)[103]. The vector autoregression was fitted in each country on about thirty years of monthly data with 14 lags. It concerned six variables corresponding not only to M, r, y and p, but also to an index (x) of the foreign exchange value of the domestic currency and to a commodity price index (pc). Recursive identification was used with the order r, x, pc, M, p, y (the order of orthogonalization did not, however, much matter because of the low contemporaneous correlations between the residuals of the autoregressive fit). Persistent negative response to interest rate innovations was systematically found for the money stock and output. Responses of the price level were less similar; the dominant pattern was an initial positive response later followed by a negative response; but in France the initial response turned out to be persistent.

5.7. The empirical approach: a discussion

1. *The VAR technique is subject to strong limitations*, which concern the various stages of the analysis[104]. Probably the main issue is double-edged and concerns, first, the question of knowing whether the VAR models provide proper characterization of the data on which they are based and, second, whether this characterization allows the interpretations given of it. The assumption of a stationary process may be a sufficient approximation in some cases or for some

[103] C. Sims, Interpreting the macroeconomic time series facts: the effects of monetary policy, *European Economic Review* (June 1992).

[104] See in particular F. Canova, Vector autoregressive models: specification, inference and forecasting, notably Section 6, in: H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics* (Blackwell, Oxford, 1995).

applications, but is clearly inappropriate in others, for instance, when we want to predict responses to a shock that has never been experienced in the past. The linear representation may look fairly innocuous to a mathematician but may be dangerous if the economist interpret it as describing the original structure of the system generating macroeconomic phenomena.

We should not forget that, in Chapter 7 and this one, we took account several times of rather visible sources of non-linearity such as capacity limitation or changes in the nature of the combination of market disequilibria. Indeed, the problem in this respect may be rather the assumption that the same matrices A_τ applied all through the sampling period. Regime changes may have occurred in the private sector and have influenced its reactions to policy interventions.

Regime changes may similarly have occurred in the determination of public policies. Neglecting this last source of change is all the more curious as there is often public direct evidence that the monetary authority itself claimed to change its policy rule or its operating procedures. When deliberately downplaying the usefulness of this evidence, VAR econometricians are taking an extreme position. Those new classical macroeconomists who opted for the opposite extreme and viewed any policy action as if it was a change in policy regime may certainly be criticized; but neglecting regime changes in an empirical study of the consequences of policy actions may also be going astray.

It is, of course, feasible in principle to apply the VAR technique for subsamples concerning periods within which mechanisms governing reactions of the private and the public sectors could be taken as fairly unchanged. But this would usually be at the cost of higher sampling variances of estimates, hence at the cost of reaching fewer statistically significant conclusions.

Assumptions embodied in the autoregressive form (219) may similarly be questioned when the analysis is applied to series that are not “detrended”. It is common practice to use seasonally adjusted series for all variables exhibiting seasonal variation. But econometricians vary in their attitude toward other prefilterings of the series analysed. Such prior treatments indeed not only remove “trends” but also introduce artificial periodicities of their own in the detrended series. However, when changes in trends or other longer-run movements are important, their presence in the analysis interferes with estimation of short-run reactions; the resulting errors have a good chance to be more important than those that would have followed from artificially introduced periodicities if the analysis had been applied to prefiltered series.

The choice of the maximum number of lags, the horizon h in (219), may also be questioned. In their analysis of monthly series Leeper et al. take a maximum lag of six months and give impulse responses over four years. In their analysis of quarterly series Christiano et al. use a maximum lag of four quarters and report impulse responses over three years. The reader of their results may be tempted to even extrapolate these results to a larger number of years. Can we be sure that a fairly good fit to statistical series on a vector autoregression with a small number of lags also provides reliable estimates of responses after a much larger number of periods? We can easily imagine examples in which such would not be the case: for instance, if a substantial part of investments would have to be replaced after three years but not before, any shock that would lead to an early investment response would also lead to another (smaller) response echoing the first one after three years; this second response would be ignored in applications relying on the validity of Equation (219) with a horizon of one year, say. Has this remark any importance in practice? It is difficult to know.

The three above remarks, about the assumed stability of the generating autoregressive process, the prior statistical treatment of economic time series, and the assumed shortness of the autoregressive horizon add to what was earlier said about the uncertainty of the identifying restrictions and perhaps still more about the uncertainty concerning the appropriate list of variables to introduce in the analysis. As a whole, they feed the fear that specification errors might endanger the reliability of the results, particularly when it is found that impulse responses or Granger non-causality tests are not robust to alternative choices of the specification.

The VAR results are presented as concerning only responses to unanticipated policy shocks. In principle they say nothing about the effects of anticipated monetary policy – or, equivalently, of the monetary policy rule – on the economy. Clearly, such a restriction of the scope of the empirical approach is uncomfortable, because there should also exist something to learn from empirical evidence about the effectiveness of policy rules. Pressed by this natural concern, economists are likely to sometimes use the VAR impulse responses as meaningful for the prediction of the effects of anticipated monetary actions. Will that be erroneous? Again, it is difficult to know.

More fundamentally, application of the VAR technique to issues such as the effectiveness of monetary policy seems to still suffer from a limitation dating from an original preconception. The initiators thought that analysis of aggregate time series alone would precisely reveal macroeconomic phenomena. They then tended to neglect all other kinds of evidence. But a large stock of empirical evidence is available beyond aggregate time series, a stock that has relevance in particular for knowing how effective is monetary policy. When this stock is examined the true role of the VAR analyses reported above better appears.

2. Indeed, *VAR analyses bring complements of, rather than substitutes for, results brought by earlier empirical approaches*. Let us consider from such a

viewpoint how complementarity may operate on the issue of monetary policy effectiveness. We must consider in turn, first, what the monetary authority does and how it reacts to the information at its disposal, second, how the rest of the economy is affected. The first part of this examination may be said to be devoted to characterization of "the central bank reaction function".

A good deal of our empirical knowledge of central banks' behaviour comes from what these banks say and explain about their decisions and about reasons why they so decided. In the ancient past this source of knowledge was the only one available. It was then used informally, and still so is most often today (the work of Friedman and Schwartz indeed contains many cases of such an informal use). But, in the US particularly, research may now draw more systematic evidence from records of the meetings of the Federal Open Market Committee, which decides on the policy of the Federal Reserve System (similar records exist in some other countries). The source was used in particular by Christina and David Romer[105]. We may simply quote them in order to present their work and summarize their results.

> "We investigated an alternative to purely statistical approaches to the question of whether monetary policy affects real economic activity. The central difficulty in answering this question is determining the direction of causation: monetary policy and various financial variables both affect and are affected by real economic developments. We argued that there is abundant nonstatistical evidence that could be extremely useful in addressing this difficulty and that economists in fact often rely on such evidence in making informal judgements about the effects of monetary policy... The goal of our paper was to investigate such nonstatistical evidence as formally and carefully as possible... Through a study of Federal Reserve records, we identified six times since World War II when the Federal Reserve appears to have in effect decided to create a recession in order to reduce the rate of inflation... We found that such shifts were consistently followed by sharp declines in real activity: 33 months after... industrial production was typically 12 percent lower than would have been predicted on the basis of real economic development up to the time of the shock."[106]

VAR econometricians could and did dispute the claim made by the Romers that, in the six identified episodes, the Fed decision to create a recession was

[105] C. Romer and D. Romer, Does monetary policy matter? A new test in the spirit of Friedman and Schwartz, *NBER Macroeconomic Annual*, No. 4 (1989).

[106] Quotation from C. Romer and D. Romer, New evidence on the monetary transmission mechanism, in: *Brookings Papers on Economic Activity*, No. 1 (1990).

relatively independent of contemporaneous real developments. These econometricians argued that concern for inflation in the Fed's meeting was a proof of the endogeneity of the policy action (see, for instance, the position taken by Sims in the general discussion of Leeper et al., op. cit.; see also a recent exchange of arguments between Leeper and the Romers[107]). But an outside observer may also think that, given the role of "plausibility" and "reasonableness" in the VAR methodology, given also the other limitations reported above, dogmatism of VAR econometricians is a bit indecent and that the work of the Romers is worth remembering: it shows that members of the Fed believed their policy would be effective and that subsequent evolution did not invalidate this belief; it adds to other reasons we too have to believe in the existence of some monetary policy effectiveness.

The minutes of the Fed's Open Market Committee are revealing in other respect also. As G. Rudebusch[108] notes, they show that, over time, the Fed's response to a given economic environment changed, that the information set on which decisions are taken is broader than recognized in VAR studies, that, on the other hand, this set does not contain statistical indicators used in such studies, but just rough proxies or at best preliminary estimates of these indicators. The most important of the three remarks probably concerns the lack of time invariance, because the other two mainly serve to remind the real complexities of phenomena which we are trying to grasp by simple and general laws.

Rudebusch (op. cit.) looks at two other kinds of evidence, which are relevant in assessing the meaning of VAR results. First, he compared VAR estimated shocks of the federal funds interest rate to another estimate of the unanticipated movements in the rate. This second estimate was based on the financial market expectations for future funds rates, as they could be inferred from the prices of federal funds futures. It did not match well the VAR shocks. For instance, a regression of the quarterly series provided by Christiano et al. (op. cit.) on the market-based series gave a R^2 of 0.20, the VAR shocks exhibiting a definitely higher volatility and the regression slope being imprecise (0.87 with a standard error amounting to 0.37).

The test is certainly valuable in the research about monetary policy effectiveness. But we may wonder whether it is as damaging for the VAR results as Rudebusch claims. It certainly suggests there is a lot of noise in the VAR shocks as estimates of unanticipated movements in the rate (even though some

[107] E. Leeper, Narrative and VAR approaches to monetary policy: common identification problems; C. Romer and D. Romer, Identification and the narrative approach: a reply to Leeper, both in *Journal of Monetary Economics* (1997) 641–665.

[108] G. Rudebusch, Do measures of monetary policy in a VAR make sense?, Federal Reserve Bank of San Francisco (January 1996).

noise in the market-based estimates for the same purpose is also likely; indeed, there may be large discrepancies between what is inferred from prices of federal funds futures and the mean anticipation of interest rates by households and firms). But no evidence of a bias in the VAR series appears, since the regression slope reported above does not significantly differ from 1. We are rather faced with a case of errors-in-variables, which may not look so surprising after the discussion of the evidence coming from the minutes of the Fed's committee meetings. The question is rather to know which bias in the estimates of the impulse responses is likely to follow from the noise affecting the VAR shocks. Let us note in passing that other Rudebusch results suggest that the noise is relatively still more important when monthly series are used in the VAR, something that will not surprise statisticians (indeed, they well know that month-to-month changes in statistical indices are erratic for a number of reasons and that quarter-to-quarter changes are likely to be more appropriate to econometric studies about business fluctuations).

Another kind of evidence might come from non-VAR econometric estimates of central banks' reaction functions. Considering such estimates of the Fed's policy rule, Rudebusch notes their instability, which he relates to the otherwise observed lack of time-invariance of the rule. But he also warns against giving weight to the evidence coming from these econometric estimates, which he considers as being embarrassingly fragile[109].

3. We now turn attention to evidence about the other part of the complex phenomenon that is at stake, namely the part concerning *the effects on the private sector*, the main part indeed regarding policy effectiveness. Here the situation radically changes, because we have so much to report as soon as we begin to look beyond the most aggregated time series. Actually the evidence is scattered throughout the economic literature, and more modestly throughout the pages of this book.

Since they began to think about the monetary policy of a central bank or government, economists thought of the cost of credit and the conditions of access to credit, for households and firms. Without any data, they already knew enough to be sure that those costs conditions sometimes mattered. The questions then were: How and how much would credit cost and conditions vary after given moves of the central bank? How and how much would those variations in turn affect supplies, demands, productions, prices...? In order to answer the questions, economists had to consider and correctly rationalize the

[109] Such an assessment would probably not be unanimously endorsed. In particular, estimates obtained by R. Clarida, J. Gali and M. Gertler, which concerned policy rules applied after 1979 and were reported in Section 4.5, did not look so fragile.

operations of the whole financial sector (deliberately neglected in this book), the behaviours of non-financial agents, finally the market adjustments. For a better knowledge of these operations, behaviours and adjustments, they defined partial models and later estimated them on the various kinds of available data. They also worried about the overall outcome and, for that purpose, built general theories and models intended to correctly portray the effects of aggregations, interdependencies and the like; later they again turned to data.

What is reported in Chapters 2 and 4 of this book is meant to give a good idea of the methods and results of the work on partial models, in particular with the role of interest rates and liquidity constraints. The subsequent chapters, up to this one and the next, similarly aim at teaching what was learned about the global phenomena, notably about the interplay between the financial system and the rest of the economy (representation of the financial system being here kept very simple). It should be clear to the reader that a large and multifarious body of empirical evidence, partly qualitative and informal, partly based on rigorous modelling and data analysis, is so ultimately serving macroeconomic knowledge.

Within this body a large part is relevant, directly or indirectly, to our understanding of the likely effects of monetary policy, and of their transmission from one variable of the economic system to another. Focusing on aggregate time series should not lead us to be blind to all other pieces of evidence, in particular to all those that contributed to our detailed knowledge of the structures and behaviour of the non-financial sector. Being more precise at this stage on the relevance of these other pieces of evidence to the issue here discussed would lead us to go too much into developments that will be better placed in the next chapter. Suffice it to say that the contribution should in particular be to clarify the basis and validity of assertions concerning the plausibility and reasonableness of VAR identifying restrictions.

But we may already hint at the nature of the other side of the complementarity, namely at the kind of contribution the VAR analyses here reported are bringing to what is otherwise learned. Very roughly speaking, we may say that direct inside knowledge of the economic world, economic modelling, analysis of microeconomic data and structural analysis of macroeconomic data lead to the best assessments we can make about the main causal links connecting economic variables and about the sizes of medium- and long-run effects, but that they leave large margins of uncertainty as to the timing of these effects, hence as to the sizes of their short-run equivalents at the horizons of three months to two years. This is explained mostly by the fact that speeds of adjustment, whether in individual behaviour or in markets, much depend on what exactly are the adjustment costs and the modes of expectations formation, two aspects about which direct knowledge or econometric evidence are uncertain. When moreover we realize that heterogeneity of conditions and behaviours has a good chance to be large in this respect, finding out the short-run time profiles of aggregate effects is particularly difficult.

Econometric analysis of quarterly aggregate time series is likely to be the best approach for the estimation of these profiles. In view of implementing the approach two methodologies have been developed, each one assuming a stochastic general model of a particular form. That was in the 1940s the "dynamic simultaneous equation model", which we shall consider in the next chapter; that was in the 1980s the vector autoregression model. The question is now to know which one of these two methodologies is best suited to the treatment of each case, or even maybe whether they should not be somehow combined. We are not yet in a position to discuss the question as regards the short-run effects of monetary policies. So let us provisionally suppose that, after examination, the VAR methodology would appear to be best suited, and let us provisionally conclude on the basis of its results (those reported here and others that the author of this book happens to know).

After a restrictive monetary policy shock, whose effect on the policy instrument (for instance, a short-term interest rate) normally extends over less than two years, the price response is difficult to characterize but certainly slow; after three years it normally does not yet amount to more than about a third, relatively to the decrease in the money stock. The output response is significant, although not large; it is usually more sustained than appears in Figure 6 and after three years not yet far below its maximum absolute value.

6 Wages, Prices and Unemployment

Surveying an important part of the empirical work on the effects of monetary policy, we just saw how explicit became the design to limit attention to the short run. Focus on a short horizon was also present from the beginning of the chapter, particularly in the models proposed in Parts 1, 2 and 4. A similar horizon will be maintained with the introduction of the labour market in the analysis, the objective being to prepare a closer scrutiny of the inflation-unemployment trade-off than was feasible up to now. But still more than earlier in the chapter a certain degree of articulation between the short and the long run will have to be recognized. Here may then be the proper place to say a word about a very common vision in applied macroeconomics, a vision inspired by what was called *the neo-classical synthesis*.

It was recently characterized by O. Blanchard[110] with two propositions: "(i) In the short run, movements in economic activity are dominated by movements in aggregate demand, (ii) over time, the economy tends to return to a steady-state growth path". The vision is obvious in the first reactions of an overwhelming majority of macroeconomists as soon as they are requested to give assessments about a concrete issue. Where they most differ is in their notion of the steady-state growth path and in their confidence about whether this notion may suffice. R. Solow seems to be more confident than others when he writes: "the trend movement is predominantly driven by the supply side

[110] Quotation taken from a session at the January 1997 meeting of the American Economic Association, Is there a core of practical macroeconomics that we should all believe?, *American Economic Review* (May 1997). The two following quotations of Solow are taken from the same source.

of the economy", and later "the demand-driven growth story sounds quite implausible to me". O. Blanchard, for instance, expresses doubts about "the nice competitive path of simple textbook models", ranks the natural rate of unemployment as a little understood feature of the real growth path and states that "macroeconomists are a long way from having a decent quantitative understanding" of growth.

The reader certainly realizes that the vision is shared by the author of this book and gives a justification to a separate study of the short run in Chapters 7, 8 and 9. However, even when accepting the short-run long-run dichotomy as a fairly reliable and convenient device for many questions, we cannot completely ignore the medium and the long run in a chapter such as this one, all the more so as we are now going to explicitly introduce wages and employment.

Discussing interactions between inflation and unemployment clearly requires consideration of the labour market. We must therefore review the set of basic notions and the first elements of formalization which served as a reference throughout Parts 1, 2 and 4; the purpose is to complement and, if need be, revise. In particular, introducing a new dimension in the analysis, we may find it appropriate to simplify in other respects. The focus on the short run will also permit a rather simple-minded representation for the labour market.

Most of our present knowledge of the interplay between inflation and unemployment directly comes from inferences using macroeconomic observations. More than others this chapter, and specially this part, thus give a high weight to results of econometric studies, and they are numerous on the subject. Typically the studies in question singled out short-run movements in prices and wages; they aimed at identifying the autonomous factors acting on one or the other, as well as at quantifying the effects of these factors.

After an introductory section which gives the background and framework, the modellization will be defined in Section 6.2. We shall then be able to turn to the simple relationships or concepts which have been popularized by the work of numerous economists, such as "cost-push inflation", the "Phillip's curve" or "the natural rate of unemployment" (Section 6.3). The following three sections will be devoted to conclusions to be drawn from available econometric studies. The brief Section 6.7 comments on the association between the dynamics of unemployment and the dynamics of inflation.

6.1. Background, objectives and framework

Within a didactic approach whose focus alternates from particular phenomena, to synthetic explanation of overall evolutions, or to policies, this Part 6 of the chapter must be seen as providing new conceptual, theoretical and empirical elements, so preparing the following part mainly devoted to diagnosis and policies.

Taking up again the formalization set up in Part 2 it will transform it somewhat with the introduction of a larger number of variables; it will so make a new step in direction of the embracing macroeconometric models to be studied in Chapter 9. We shall reflect in a moment on the background of this new step. But in order to put the reflection in a proper perspective, we must realize that complements have to be brought also to the conceptual framework within which models will be used, we have to briefly consider how introduction of wages and employment relates to what was discussed up to now, earlier chapters included, and we have to remember that the theory aims at being relevant to a wide range of situations, in particular to those in which inflation is acute and may require more than the familiar fiscal and monetary policies.

Some basic concepts for the short-run analysis of the labour market were presented in Chapter 7, particularly in Section 3.1, but only those concepts required for the static analysis that was then at stake. Now, as we are dealing with inflation, we shall soon meet the NAIRU, the "non-accelerating-inflation rate of unemployment". Its presence requires a reconsideration of our conceptual system, a reconsideration that will involve more than one or two new distinctions. In the same way as the Beveridge curve provided an empirical anchor in the definition of frictional unemployment, the Phillips curve will make the NAIRU operational for applied analysis. But the empirical anchor and the operationality of the concept will exist only subject to certain conditions, which we shall examine. So doing we shall look at the inflationary process from a somewhat new interesting angle, but really recall much older beliefs in the existence of links between changes in wages and the degree of the labour market slack or tension. Our models will have to be suited to incorporation of such beliefs.

Introducing wages in the analysis we are now able to take into account the fact that the mutual relationships between the pressure of demand and inflation are affected by induced changes in the distribution of income, especially in the short run. This effect was ignored in the previous parts where attention often focused on a different but somewhat related distinction, that between anticipated and unanticipated changes in real income.

Earlier in this book we have already considered the impact on aggregate demand of changes in income distribution: in Section 2.8 of Chapter 2 when we summarized factors which might explain aggregate household saving, in Section 2.1 of Chapter 6 for the introduction to some theories of growth and in Part 4 of Chapter 7 in which the impact in question played a part in explaining the different comparative statics properties of Keynesian and classical unemployments. The impact comes from the fact that different categories of income are not subject to the same propensity to save, or rather that different categories of income are spent at different rates and therefore are found sooner, or later, in consumption or in investment. An increase in low wages relative to the income of the self-employed stimulates demand, at least to begin with; an increase in corporate profits relative to household incomes usually slows down demand, since company spending programmes are more bulky and less easily revised than those of households.

Thus, the complications we are going to incorporate in this part will naturally fit in the study of economic policies acting through their impact on aggregate demand. The logic will be absolutely the same as the one studied so far in this chapter and in the previous one. But new considerations will also emerge when attention will focus on cases when inflation is rapid. Indeed, inflation then consists of a more or less anarchic escalation of different prices, of different wage rates, of different rates of return on capital. Public policy is naturally called to directly intervene in this escalation. We then speak of a price and income policy. Our macroeconomic models ought to be suited also to the study of the macroeconomic aspects of such a policy.

In concrete situations it is often easy to identify some "shocks" that have hastened the rise in prices and which have had widespread repercussions on the whole set of nominal values. These shocks vary greatly from one situation to another: a bad harvest which makes staple foods more expensive, a policy to raise low wages, a strengthening of monopoly positions on markets for important basic commodities, etc.

To react lucidly against inflation policy makers at times try to avoid or to lessen the shocks, when they think they have the power to do so; otherwise they try to mitigate the effects. Measures to prevent the initial shocks involve structural policies, which we are not concerned about (promoting a productive agriculture less exposed to risks, compensating for the increase in low wages by stabilizing higher wages, implementing a competition policy, etc.). On the other hand, we ought to throw some light on the likely consequences of those measures which aim either at preventing the transmission of nominal increases or even more at containing an overall acceleration of inflation. For this, we have to understand how an increase affecting a price or a remuneration rate tends to be transmitted to other prices or to other rates of return. This is why we are going to consider in this part the main vehicles of transmission, that is, the effect of price increases on wages and the effect of wage increases on prices.

So, whether it is a question of looking for the conditions under which some common concepts are valid, of completing the analysis of the link between the pressure of demand and inflation, or of setting up a framework for studying price and income policies, we must enrich the models of the previous parts by introducing a simple formal representation of how the labour market functions in the short run. So as to avoid analytical complications we shall not directly build on to the model defined in Part 2. It is, however, interesting to examine briefly what would result if we should follow this direct path; indeed, in applied models which one may want to construct for economic forecasting, one may find it appropriate to maintain distinctions that we shall ignore in our subsequent theoretical study.

To represent equilibrium on the market for goods and services we had, in Part 2, neglected the random disturbances affecting the adjustment process and the formation of demand. We shall do the same in this part, except marginally when reporting econometric results. The adjustment process involves four essential variables: the rate of price increase g_p, the expected rate of price increase g_p^e, the degree of utilization of productive capacity q and the pressure of the demand for goods d. Since we want to consider the labour market as well, we should in principle, introduce four new variables: the rate of wage increase g_w, the expected rate of wage increase g_w^e, the rate of employment of the labour force, or equivalently the rate of unemployment, and the pressure of the demand for labour, which could be measured as the ratio between job vacancies and unemployment.

The model of Part 2 includes four main relationships for determining the four main variables:

– how expectations about price inflation g_p^e are formed in terms of past observations (see Equation (4)),
– the formation of the pressure of the demand for goods d (Equation (58)),
– two adjustment laws specifying how the degree of capacity utilization q and the rate of inflation g_p evolve (Equations (34) and (35)).

If we should transpose directly, we should have to introduce four additional main relationships concerning:

– expectations as to the rate of wage increase,
– the pressure of demand for labour,
– two adjustment laws specifying evolution of the rate of unemployment and of the wage rate.

These new relationships would obviously have to be linked to the previous ones: the pressure of the demand for labour depends on the demand for goods; evolution of the wage rate is affected by expectations about price inflation, etc. The presentation and study of a model organized around these eight main relationships could be of interest. But it would be laborious, as we may imagine after what we had to do in Parts 2 and 4. So we are going to be satisfied with a simpler version of the model, a version which will, however, suffice for the discussion of now widespread ideas, also as a framework for relevant econometric work, and still for the study of major problems posed by policies intended to curb inflation.

6.2. Modellization

1. The way to *simplify the model* we have just sketched is obvious. First of all we must reduce the number of variables, and to do that, we want to have just one quantity variable to reflect the degree of imbalance on each of the markets. We must then define a single variable to represent both the degree of utilization of the resource (productive capacity or labour force) and the pressure of demand (for goods or labour).

This simplification is not drastic if we assume, as we did on occasion previously, that there is, in effect, a perfectly stable functional link between changes in the degree of utilization and changes in the pressure of demand. Equation (36) which expressed this link for the goods market did not have to include either expectations as to future inflation, or economic policy measures; it was supposed to show the only feature that mattered, namely that adjustment costs associated with changes in the rate of production lead to a gradual adaptation in capacity utilization.

Given this, to simplify the presentation of what follows, we are going to keep the notation d_t and continue to talk about the pressure of demand for goods; but we are going to modify the definition by substituting for (32):

$$d_t = \frac{D_t}{\bar{y}} - \hat{q} \tag{227}$$

(thus d_t replaces what was $d_t + q_t - \hat{q}$ in Part 2). So, this is a relative measure of excess demand above that level which would correspond to a normal degree of capacity utilization. The ratio of capacity utilization itself q_t no longer features explicitly. Note in passing that introducing in Part 2 only d_t, with this new definition, would have led us to posing the following equation for the rate of inflation g_{pt}:

$$g_{pt} = g_{pt}^e + \pi(d_t) \tag{228}$$

with a decreasing function π. Equation (34) would, in fact, have had this form if we had assumed that $k_3 = k_4$, whereas we preferred to keep the inequality $k_3 > k_4$.

Where labour is concerned, we shall proceed in the opposite manner so as to conform to standard practice, to show explicitly a variable which has intrinsic interest. So we shall not introduce into the model the pressure of the demand for labour, but on the contrary the rate of unemployment denoted u_t. Implicitly, we shall assume that there is a perfectly stable functional link between changes in the rate of unemployment and changes in the pressure of the demand for labour. We should keep in mind, however, the fact that the rate of unemployment adapts gradually (that is with some delay) to the changes in the pressure of the demand for labour, since hiring and firing are seldom instantaneous.

2. Actually, in the short run, the two variables d_t and in u_t, which reflect the situation on each of the two markets, are highly correlated. *Strong pressure on the goods market* (d_t *high*) *usually corresponds to strong pressure on the labour market* (u_t *low*). We shall take advantage of this correlation to simplify even further the structure of our model by introducing a functional link between changes in d_t and in u_t. This means that we shall not deal with questions which would involve divergent effects on the goods market and on the labour market. In other words, we are not going to try and make the distinctions of Chapter 7, Part 4, dynamic, nor to adapt them to the study of inflation. However, in applications we may have to consider several possibilities as to the context in which inflation will be studied: depending upon the case, the medium-term level of the rate of unemployment will be higher or lower, and this may have repercussions for some of the equations which will be used.

Be that as it may, we shall assume here that u_t reacts with lags to the changes in d_t; we shall thus set:

$$u_t = E(L)d_t + e, \tag{229}$$

where $E(L)$ is a rational function of the lag operator L. In other words (229) stands for the following equation:

$$u_t = \sum_{\tau=0}^{\infty} e_\tau d_{t-\tau} + e, \tag{230}$$

where $E(L)d_t$ is assumed to have a convergent power series expansion in τ.

An expression like this, where e is a positive constant and the e_τ are non-positive constant coefficients, is certainly acceptable as a first approximation for normal times. The assumed (somewhat lagged) adjustment of unemployment to changes in the demand for goods makes sense when the behaviours of employers and employees is considered. We just have to think of three successive adjustments: firms adapt their output to the demand addressed to them; firms adapt their demand for labour and their employment to their output; households adapt their supply of labour to the state of the labour market, hence to their expectation of seeing this supply fulfilled. In Chapter 4, Section 1.4, we focused on the second adjustment and pointed to the empirical evidence supporting an equation where employment was determined from output through a distributed lag formula. In Chapter 2, Section 3.6, we paid attention to the third adjustment when speaking about the concept of "disguised" or "hidden" unemployment.

We may also quote here a well established empirical regularity called *the Okun law*, covering the joint effect of the second and third adjustments: an increase in the level of output leads to a proportional but smaller decrease in unemployment (A. Okun said -0.3 per cent in u_t for $+1.0$ per cent in y_t). Defining the Okun law almost exactly as relation (229) with replacement of d_t by the ratio between output y_t and its level $\bar{y}_t$ at a normal degree of capacity utilization, R. Gordon estimates the sum of the e_τ at -0.5 for "most of the postwar period in the USA"[111].

We could, of course, refine the analysis of the three kinds of adjustment. This would lead to a less simple outcome than Equation (230). As we hinted above, the conditions of supply and demand on both the goods market and the labour market would then play a part. Rather than making the analysis more complex, we may stick to Equation (230) and remember that it is meant to apply within a period during which market condition will not radically change. We must then accept that the values given to e_τ and e might depend on the type of market situation perceived as holding over several years. If we are considering a country and a period when there is little unemployment, we accept the idea that the absolute values of e_τ and of e should be small; the demand for labour is then likely to be rationed on most labour markets (by profession and region); unemployment should not be much sensitive to changes in demand. If, on the other hand, unemployment affects almost all professions and regions, the values of the e_τ should be almost independent of the medium-run level of this unemployment, whereas the value of e would obviously depend on it.

3. We now have to specify *how are determined the growth rate g_{pt} of the overall price level and the growth rate g_{wt} of nominal wage* rates. In contrast

[111] R. Gordon, The time-varying NAIRU and its implications for economic policy, *Journal of Economic Perspectives* (Winter 1997), footnote 4, p. 15.

to equations such as (34) or (228) we are going to model things in a slightly different way, in the sense that we now want to be explicit about the different causes of price and wage increases and this will lead to our no longer making expectations prominent. Indeed, recall how, in Section 2.2, we justified the idea that the actual price increase depended on the expected increase: we said that producers took into account their expectations about wages, and wage earners took into account their expectations about prices. We must now make this reciprocal interdependence between to two explicit, and its rationale involves various adjustment costs, which may play a more important role than expectations *stricto sensu*. Moreover, whenever expectations depend on previously observed values as in Equation (4), we do not need to explicitly bring them in if past values are already taken into account. (This remark shows that an exceptional change in expectations should then be counted as an exogenous cause of changes in prices and wages and be included in the measure of z_{pt} or z_{wt}; this will be assumed to be done in applications of the following model.)

We write the price adjustment law as:

$$P_p(L)g_{pt} = P_w(L)g_{wt} + z_{pt} + \pi(d_t). \tag{231}$$

Three terms appear in the right-hand member:

- the first $P_w(L)g_{wt}$ expresses the influence of past or contemporaneous wage increases, whether these act directly, or through wage expectations,
- the second z_{pt} captures all exogenous causes of price changes: for example, changes in indirect taxation and, for an open economy, changes in the price of imported raw materials,
- the third $\pi(d_t)$ expresses the direct effect of the pressure of demand for goods on prices.

In the left-hand member the rational function $P_p(L)$ characterizes the lags in price adjustments, in particular those due to price expectations. The first of the coefficients $p_{p\tau}$ in the polynomial, namely p_{p0}, is assumed equal to 1.

We can write the wage adjustment law symmetrically, as:

$$W_w(L)g_{wt} = W_p(L)g_{pt} + z_{wt} + \sigma(u_t). \tag{232}$$

The second and third terms in the right-hand member are easy to interpret. One concerns the influence of all exogenous causes acting on wages (increases in the minimum legal wage, exceptional results of national wage agreements, exogenous changes in relative bargaining strength in labour disputes, etc.); the other expresses the idea that the higher the pressure of demand for labour, the faster the increase in wages (σ is thus a decreasing function of u_t). The first term $W_p(L)g_{pt}$ shows the influence that increases in the price level have on nominal wages: those increases affect the purchasing power of wage earners income as well as the value of sales, and hence the financial capacity of firms. It is therefore natural to expect that the faster prices rise, the faster wages rise. Finally we assume $w_{w0} = 1$.

Equations (231) and (232) result from obvious simplifications. The most important is to pretend to capture with just two aggregate variables z_{pt} and z_{wt} all those effects which do stem neither from the levels of prices and wages nor from the pressure of demand on the goods and labour markets. In applications we shall usually have to express each of these variables as a sum of terms to be more narrowly and precisely defined. It may also be necessary to recognize that some of these effects are, in part, induced by the inflationary process and should not, therefore, be treated as purely exogenous.

We should also note that there are different ways of applying and interpreting equations such as (231) and (232). If we want to concentrate on the very short run, the variables z_{pt} and z_{wt} can capture long delayed impacts of former inflation, such as late alignment of public tariffs after cost rises, or even late increase in the price of some imports after a sharp depreciation in the exchange rate. If on the other hand we want to deal with longer-run phenomena, the same effects should be captured within the terms $P_p(L)g_{pt}$, $P_w(L)g_{wt}$, $W_w(L)g_{wt}$ and $W_p(L)g_{pt}$ rather than by z_{pt} and z_{wt}. This explains why the specification of the functions $P_p(L)$, $P_w(L)$, $W_w(L)$ and $W_p(L)$ may depend considerably on the applications we have in mind; in particular the values of $P_w(1)$ and $W_p(1)$ characterizing the importance of what we may call the direct "long-run" effects of wages on prices and of prices on wages heavily depend on the application, since what we mean by the long run varies according to the question we are dealing with.

But, even more than Equation (229), Equations (231) and (232) also depend on the real context considered. The specification to choose for $W_p(L)$, in particular, is not at all the same depending upon whether we are considering an economy which is not very sensitive to inflation, or on the contrary, an economy where there is a complete indexation of wages to prices. This is one point that econometric estimations bring out very clearly. Likewise, changes in the institutional context may react on the importance of the effects of the pressure of demand d_t and of the rate of unemployment u_t. The more the economy is regulated and controlled, or governed by long-term contracts between suppliers and customers, between employers and employees, the less important is the part played by the law of supply and demand, and the weaker are the effects of d_t and u_t. Likewise again, we should be aware that the same rate of unemployment may not have the same effect on the formation of wages in a situation where the labour force is almost exclusively masculine and adult as it does in another situation where a substantial proportion of people looking for work might consider the possibility of remaining inactive. We should also take into account the fact that better unemployment benefits reduce the urgency of finding a job and mean that people looking for work do not have to accept a salary when it appears to them to be low. In other words, the function $\sigma(u_t)$ depends on particular circumstances.

A fortiori, if we want to linearize the functions $\pi(d_t)$ and $\sigma(u_t)$, the model should be considered as dependent on a given medium-run context. The sensitivity of g_{pt} to d_t and that of g_{wt} to u_t turns out to be low in a depressed economy, where the pressure of demand is low on the goods market as well as on the labour market. Each of these terms plays a more important part when we pass from a situation where the market in question is depressed to another where it experiences excess demand.

4. The two equations (231) and (232) are meant to serve in the dynamic study *of short-run overall movements of prices and wages*. They derive from direct and rather simple ideas on price and wage adjustments. They will underlie the econometric specifications and results to be discussed later. They will enter into our further exploration of anti-inflation policies. Let us examine them a little more closely and consider various ways of using them.

First, although concerned here with short-run dynamics, we must wonder how the equations would fare if we would rely on them for longer-run analysis. We should, indeed, be worried if they would make no sense at all when viewed from such a broader perspective. Simple tests are based on the comparative statics properties of stationary solutions to the system made of these equations, and on the stability of such solutions.

This is a partial system for the simultaneous determination of the series of the two rates g_{pt} and g_{wt}. A stationary solution must give constant values g_p and g_w to these endogenous variables and requires constant values z_p, z_w, d and u of the exogenous variables. Then (231) and (232) translate into two equations linking the six constant values in question, namely:

$$P_p(1)g_p = P_w(1)g_w + z_p + \pi(d), \tag{233}$$

$$W_w(1)g_w = W_p(1)g_p + z_w + \sigma(u). \tag{234}$$

Clearly, we should like this system to have one and only one solution whatever the values given to the exogenous variables[112]. This requires that:

$$P_p(1)W_w(1) - P_w(1)W_p(1) \tag{235}$$

be different from zero.

We may assume this condition on the four rational functions of L to hold. It holds in particular for focal cases in which:

$$P_p(1) > P_w(1) > 0, \qquad W_w(1) = W_p(1) > 0. \tag{236}$$

The equation means exact long-run indexation of wage inflation on price inflation, because a change of one percent in g_p would imply according to (234) an equal change in g_w, for given values of z_w and u. In contrast the left inequality in (236) implies less than full long-run indexation of prices on wages in (233); in other words, if z_w and u do not change but there is a permanent increase in z_p or d, the long-run rate of inflation, measured by either g_p or g_w, will increase. Perhaps such a property would seem foreign to some theories of the long run path of a monetary economy; but it would then probably be because such a theory would not admit the possibility of permanent changes in concepts such as z_p or d, which assume the existence of some form of market disequilibrium. The property indeed makes sense when we think about what may happen during a decade or more within a phase of a long wave.

We shall not much discuss the stability of solutions to the system (231)–(232); aiming at a full discussion would take much space, although it ought just to apply the mathematical theory of linear dynamic systems with constant coefficients. For instance, assuming stationary values of z_{pt}, z_{wt}, d_t and u_t, convergence to the stationary solution would hold for all sets of initial values if and only if the $2h$ solutions λ_j of

$$\det\left[\sum_{\tau=0}^{h} A_\tau \lambda^\tau\right] = 0 \tag{237}$$

would have moduli larger than 1, the 2×2 matrix A_τ being that of coefficients of λ^τ in

$$\begin{bmatrix} P_p(L) & -P_w(L) \\ -W_p(L) & W_w(L) \end{bmatrix}.$$

[112] Actually it makes sense to say that these values cannot be arbitrarily chosen because $g_w - g_p$, the long-run growth rate of real wages, is bound to be equal to that of labour productivity. But imposing the value of $g_w - g_p$ would not interfere with the argument given here.

For instance, in the case when $h = 1$, when A_0 is the identity matrix and A_1 is non-positive (there is no contemporaneous effect of prices on wages or of wages on prices; lagged effects are non-negative) and when the determinant (235) is positive, stability holds for all choices of initial values.

At this point an incidental remark is in order, because it leads to consideration of an alternative wage equation to (232). We noted that the equation in (236) implies full long-run indexation of wage inflation on price inflation for given values of z_w and u. But it does not imply full long-run indexation of wages themselves on prices. Moreover, a permanent change in the constant value z_w or u implies a permanent change in $g_w - g_p$, hence an increasing discrepancy of the real wage from its earlier course. These features do not fit well with heuristic ideas about the wage dynamics we are here discussing.

This is why econometricians and macroeconomists, concerned with the long-run properties embodied in their specifications, often now introduce an "*error-correction*" component in the wage equation. It amounts to adding in the right-hand member of (232) an extra term, most often of the form:

$$W_v(L)\big[\log(w_t/p_t) - \hat{v} - \gamma t\big], \tag{238}$$

where $\hat{v}$ and γ are exogenous constants and where $W_v(1) < 0$.

The presence of this term has the effect of attracting the real wage toward an exogenously given path, here assumed to be an exponential growth at rate γ. A stationary solution of the system so amended requires that the square bracket in (238) has a constant value, hence that $g_w - g_p = \gamma$. Changes in exogenous constants, such as z_w and u but not γ, only affect the stationary value of the square bracket, no longer the discrepancy between the two rates of inflation.

We note in passing that, since $\sigma(u)$ is a decreasing function and $W_v(1) < 0$, the constant value of the real wage w/p in a stationary solution is a decreasing function of the unemployment rate u. The property agrees with empirical observations gathered by D. Blanchflower and A. Oswald under the name of "*the wage curve*" (see the end of Section 2.4 of Chapter 4).

Clearly, altering (232) by introduction of an error-correction term would lead to a more accurate wage equation, but would also much complicate the subsequent analysis. For simplicity we shall stick to (232) as written, hoping that the error correction will not in fact play an important part at the horizon we shall have in mind.

5. Let us now consider *the dynamic system* made of (229), (231) and (232). There is a natural way to read it, namely to view the series of d_t, z_{pt} and z_{wt} as exogenous and the system as determining the series of the endogenous variables u_t, g_{pt} and g_{wt}. Solution of the system gives in particular the series

of the rate of price inflation. Focusing on this rate, and on how it depends on the past and contemporaneous values of the pressure of demand, would very much remain in the spirit of large parts of the chapter up to now, with only the explicit addition of other factors than pressure of demand in the explanation of inflation.

With this approach it is natural to use (229) and (232) as giving the series $\{u_t\}$ and $\{g_{wt}\}$ as functions of the series $\{d_t\}$, $\{g_{pt}\}$ and $\{z_{wt}\}$, and then to introduce those functions in (231), so eliminating $\{u_t\}$ and $\{g_{wt}\}$. The result of such an operation is a "reduced" dynamic equation on the price-inflation rate (for the time being, we assume that an inverse to the determinant $P_p(L)W_w(L) - P_w(L)W_p(L)$ exists and is well behaved, a point on which we shall come back in the next section). Assuming the functions π and σ to have been approximated by linear functions for the purpose, the dynamic equation is linear. We may write it here as:

$$g_{pt} = \widehat{B}_p(L)g_{p,t-1} + \widehat{P}_p(L)z_{pt} + \widehat{P}_w(L)z_{wt} + \widehat{P}_d(L)d_t + \hat{a}. \tag{239}$$

An equation of this type is sometimes used in applied macroeconomics. But special attention being given to the NAIRU since four decades, an alternative specification for the reduced dynamic equation is more often considered, one in which tensions on the labour market is given pre-eminence over tensions on the market for goods and services. The equation then writes somewhat like the following:

$$g_{pt} = \widetilde{B}_p(L)g_{p,t-1} + \widetilde{P}_p(L)z_{pt} + \widetilde{P}_w(L)z_{wt} + \widetilde{P}_u(L)u_t + \tilde{a}. \tag{240}$$

It is a natural question to ask which of the two formulations (239) or (240) should be preferred. On the one hand, we may say that the second vividly exhibits the main dilemma faced by macroeconomic policy, namely whether to give priority to the concern for inflation or to the concern for unemployment. Equation (240) directly gives the trade-off. We shall also see, in the next section, how in the 1960s the empirical approach to the study of short-run macroeconomic phenomena naturally led to consider such an equation.

On the other hand, the causal interpretation is less indirect, hence more transparent, with (239) than with (240). Both are reduced equations, but elimination of $\{u_t\}$ and $\{g_{wt}\}$ obscures less the causal links underlying the "structural" system made of (229), (231) and (232) than does elimination of $\{d_t\}$ and

$\{g_{wt}\}$. Indeed, the pressure of demand d_t naturally appears as exogenous in the structural model in question, which is however a partial model; Equation (229) is indeed meant to explain unemployment u_t from contemporaneous and past values of d_t: the intensity of slack on the labour market is seen as due to current and earlier degrees of slack on the goods market. In order to obtain Equation (240) we have to consider not directly (229) but the inverse relation, which we may write:

$$d_t = \left[E(L)\right]^{-1}(u_t - e). \tag{241}$$

But this cannot be interpreted as a causal relation. Thus, we cannot in this case explain the causal meaning of the reduced equation without speaking about the structural system. The short cut involved in side-stepping the loop between wage and price inflations, whether for (239) or for (240), is in contrast transparent.

Actually writing (241) is tantamount to assuming that an autoregressive representation of d_t in terms of u_t can be found, in other words assuming that the rational function $E(x)^{-1}$ has a convergent power expansion (remember what was said when the operator L was introduced). This is not the case for all conceivable meaningful specifications of $E(L)$. But we shall forget about the possible difficulty.

Equations (239), or (240), still involves values of $g_{p,t-\tau}$ (for $\tau > 0$) on the right-hand side. It is of course possible, and indeed convenient for further calculation, to fully solve for g_{pt} in terms of other variables. This is easily done thanks to a simple multiplication of the equation by $[I - \widehat{B}_p(L)]^{-1}$, or $[I - \widetilde{B}_p(L)]^{-1}$. But the form (239), or (240), is closer to what is most often found in econometric results, to which we shall turn attention after the next section.

Let us, however, consider the following form, which will serve for a discussion of the NAIRU:

$$g_{pt} = P_p^*(L)z_{pt} + P_w^*(L)z_{wt} + P_u^*(L)u_t + a^* \tag{242}$$

and is derived from (240) after multiplication by $[I - \widetilde{B}_p(L)]^{-1}$. In order to get a feeling about the type of dynamics implied by the adjustment model developed up to this point, a particular specification and even a numerical example will be useful.

As for the specification of the interplay between price and wage inflations, let us take a case in which (236) applies:

$$P_p(L) = W_w(L) = 1, \qquad P_w(L) = \frac{c(1-b)}{1-bL}, \qquad W_p(L) = \frac{1-b}{1-bL} \tag{243}$$

with

$$0 < b < 1, \qquad 0 < c < 1.$$

Solving for g_{pt} the system (231)–(232) we obtain:

$$\begin{aligned} g_{pt} = {} & (1-bL)^2 D(L)\big[z_{pt} + \pi(d_t)\big] \\ & + c(1-b)(1-bL)D(L)\big[z_{wt} + \sigma(u_t)\big] \end{aligned} \tag{244}$$

with

$$D(L) = \big[(1-bL)^2 - c(1-b)^2\big]^{-1}. \tag{245}$$

Moreover, linearizing the two functions $\pi(d)$ and $\sigma(u)$, and specifying $E(L)$ in (229), we write:

$$\pi(d_t) = \pi_0 + \pi d_t, \qquad \sigma(u_t) = \sigma_0 - \sigma u_t, \tag{246}$$

$$E(L) = -\eta(1 + \alpha L), \tag{247}$$

π, σ, η and α being positive numbers. Introducing these expressions in (244), we find:

$$P_p^*(L) = (1-bL)^2 D(L), \qquad P_w^*(L) = c(1-b)(1-bL)D(L), \tag{248}$$

$$P_u^*(L) = -m(1-bL)D(L)\frac{1+nL}{1+\alpha L} \tag{249}$$

with:

$$m = c(1-b)\sigma + \frac{\pi}{\eta}, \qquad mn = c(1-b)\alpha\sigma - b\frac{\pi}{\eta}. \tag{250}$$

As a numerical example let us pose:

$$b = \alpha = \frac{3}{4}, \qquad c = \frac{1}{2}, \qquad \pi = \sigma = 1, \qquad \eta = \frac{4}{7}, \tag{251}$$

which imply $m = 15/8$ and $n = -0.65$. The long-term effects of permanent changes in z_{pt}, z_{wt} or u_t are then given by:

$$P_p^*(1) = 2, \qquad P_w^*(1) = 1, \qquad P_u^*(1) = -3. \tag{252}$$

The corresponding contemporaneous effects are respectively 1.03, 0.13 or −1.94. After the first four "quarters" ($0 \leqslant \tau \leqslant 3$) the cumulated effects are 1.49, 0.39 or −1.30; after the first five quarters they are 1.62, 0.45 or −2.00 (the denominator in the right-hand side of (249) implies that the impulses from a given temporary shock on the unemployment rate alternates between positive and negative in subsequent quarters, because return of u_t to its former value and absence of any subsequent change in this variable requires alternating positive and negative values of the demand pressure d_t). Although coming from a simple-minded example, these results exhibit the kind of short-run effects we can expect, particularly when we recall that in fact seasonally adjusted series of the unemployment rate are fairly smooth.

6.3. The NAIRU, the natural rate of unemployment and a few other familiar concepts

Consideration of the labour market in the macroeconomics of inflation led to the introduction of concepts which, although widely used since several decades, are not uniformly defined in the same way and are therefore source of ambiguity, although quite suggestive. We are now equipped to examine these concepts and to relate the theory of inflation to the usages most frequently found in the literature. We shall first briefly come back to the distinction, introduced in Section 1.2, between cost-push and demand-pull inflation; we shall then discuss the concept of NAIRU; this will give us the reference for looking back at the developments which followed the introduction of the "Phillips curve"; we shall end with the concept of natural rate of unemployment.

1. *We often set cost-push inflation against demand-pull inflation*. By this we mean that, in the first case, the origin of price increases lies in an exogenous increase in production costs, whereas in the second case, it is the result of an excessive pressure of demand on various markets. This is a very convenient

terminological distinction, even though it is sometimes difficult to apply objectively, as we saw in Section 1.2. It is in the very nature of the inflationary process to have complicated links between costs, prices and demands. To draw conclusions about this process, or about its acceleration, we have to assume that we can distinguish between the initial causes and the resulting effects, something which may not be obvious (even the increases in the price of oil in 1973 and in 1979, which are usually considered as having accelerated cost-push inflation, have sometimes been attributed to demand pull).

Equations (231) and (232) give a more precise meaning to these commonly used expressions. We have to speak of cost-push inflation to indicate the changes which accompany and follow an increase in the variables z_{pt} and z_{wt}. We have, in the same way, to speak of demand-pull inflation, if autonomous causes were responsible for an abnormally high value of d_t or an abnormally low value of u_t.

These equations, modelling the inflationary process, directly make the distinction in a more precise setting than Equation (2); they lead us to pinpoint where the origin of a change in the process is identified as occurring. Room remains for discussion since, as we realized in the preceding section, the utilization and interpretation of the model in a particular case can be flexible, the same historical time paths being analysed from different points of view. But it makes sense to say that, in Western Europe, cost-push inflation accounted for the 1974 increase in prices, whereas demand-pull had formerly characterized the slow but persistent acceleration of price increases since the late 1960s.

A regrettable practice was established in the economic language in the 1970s, namely to speak not only of demand shocks as being responsible for demand-pull inflation, which is all right, but also of "supply shocks" as being responsible for all kinds of cost-push inflation, and this introduced a confusion. A change in z_{pt} or z_{wt} may be due to something that can validly be called a supply shock, but it may also be due to an autonomous shock in wage negotiations, or in the legal minimum wage, or in the terms of foreign trade, or still in indirect taxes. If we want a general label for this last kind of shocks, we should rather use an expression such as "adjustment shocks" or "autonomous wage–price shocks".

The distinction between the three kinds of shocks may be clarified by reference to the macroeconomic theory of market disequilibria. If we consider the categories discussed in Part 4 of Chapter 7, more particularly those introduced in Section 4.2, we may say that a demand shock bears on the consumption function f, on investment I or on public demand G; a supply shock bears on the production function F or the labour supply function h; an autonomous price–wage shock bears on the price level p, the wage rate w or the tax rate τ. The link between these various shocks and the course of inflation was not systematically made there, neither in the purely static theory which was then exposed, nor in the dynamic models of Sections 5.4 and 5.5, where the laws assumed for changes in prices and wage rates were not much discussed. With other simplifications the present chapter offers a complementary hindsight, which will however not be elaborated on here.

How variables z_{pt} and z_{wt} can capture autonomous price or wage shifts is obvious. But they can account also for true supply shocks. Let us consider, for instance, a decrease in the productivity growth rate. It cannot be introduced as such in the simple dynamic model made of Equations (229), (231) and (232) since productivity does not appear. If anything, the model may be read as implicitly assuming a constant rate of productivity growth. With respect to this reference, the actual decrease in productivity growth plays like a proportional increase in z_{pt}, the only difference is that we must be more careful than we have been so far. In other words, we must specify how the model had to be used for an economy with a constant rate of productivity growth: the reference path growth was for instance one in which $\pi(d_t) = \sigma(u_t) = z_{wt} = 0$, with moreover $g_{pt} = 0$ but with a positive g_{wt} and a negative z_{pt}, being so calibrated as to be consistent with the growth of productivity. The negative productivity-growth shock, now represented by an increase in z_{pt}, will result in a positive g_{pt} and, all effects being taken into account, in a lower rate of growth $g_{wt} - g_{pt}$ of the real wage.

2. The "*non-accelerating-inflation rate of unemployment*", the NAIRU, explicitly refers to inflation and is indeed a more important concept for the analysis of inflation than for that of unemployment.

The definition takes its full meaning with reference to the price dynamics exposed in the preceding section. It may be phrased as follows: "the unemployment rate that is consistent with steady inflation in the absence of transitory supply or price–wage shocks"[113]. According to a first interpretation of the definition, the NAIRU is directly read from Equation (242), once constant reference values have been given to the other variables than u_t, namely $g_{pt} = \hat{g}_p$, $z_{pt} = \hat{z}_p$, $z_{wt} = \hat{z}_w$, the resulting constant value $\hat{u}$ measures the NAIRU: starting from there and keeping z_{pt} and z_{wt} fixed at $\hat{z}_p$ and $\hat{z}_w$, a decrease in u_t would lead to a progressive increase in g_{pt}, i.e. to an acceleration in the price increase.

[113] This is almost exactly the definition given, for instance, in R. Gordon, The time-varying NAIRU and its implications for economic policy, *Journal of Economic Perspectives* (Winter 1997) 15. Gordon does not mention in his definition that permanent supply or price–wage shocks should affect the definition of the NAIRU.

With this interpretation the NAIRU clearly appears as a benchmark for the study of demand-pull inflation and of policies intended to control it. (Since this control requires time, it cannot compensate for the impact of transitory shocks; hence their exclusion in the definition.) Like other benchmarks its exact level is conventional, since it depends on the reference values given to other variables. The best convention may change when one moves from one country to another, from one period to another distant period. All the more so as estimation of Equation (242), meant to characterize short-run movements, may change: not all relevant variables have been identified in z_{pt} and z_{wt}, the adjustments in the economy may have become more or less complete, and so on.

There is nothing fundamentally wrong in the statement that the level of the NAIRU might depend on a reference rate of inflation $\hat{g}_p$. Indeed, such a reference is very much present in the mind of policy makers. For instance, they may think in terms of a target inflation rate like 3 percent a year. Also the value of the NAIRU may not much vary with changes in $\hat{g}_p$ because $P_u^*(1)$ may be large in absolute value: we found 3 in our numerical example, and it would have been easy to find much more, for instance, with a value of the parameter c larger than 0.5, perhaps approaching 1.

However, it may be a little disturbing to wonder why the definition of a concept attached to the notion of "non-accelerating-inflation" ought to be linked to an equation meant to explain the rate of inflation and not its change $g_{pt} - g_{p,t-1}$. According to (242) this change is explained by contemporaneous or lagged values of $u_t - u_{t-1}$, not of u_t itself. There is, indeed, a second interpretation of the definition, an interpretation which, however, requires a marginally different and more narrow specification of the dynamics of inflation.

3. In route through our derivation of (242) we assumed in passing that the rational function of L given by the determinant $P_p(L)W_w(L) - P_w(L)W_p(L)$ had a well behaved inverse, i.e. one that, first, existed and, second, allowed convergent power expansions. The condition is met in most econometrically fitted models. But there are theoretically interesting, and even empirical, models in which such is not the case.

For an introduction let us consider the very simple example of the system:

$$g_{pt} = g_{wt}, \quad g_{wt} = g_{p,t-1} - \sigma\big(u_t - \hat{u}\big). \tag{253}$$

There would be complete and immediate indexation of prices on wages, like with complete and universal cost-plus pricing; the indexation of wages would have a lag of one period and the growth of the real wage would depend on the unemployment rate. This system implies

$$g_{pt} - g_{p,t-1} = -\sigma\left(u_t - \hat{u}\right). \tag{254}$$

A constant rate of inflation requires the unemployment rate u_t to be equal to the NAIRU, which remains the same, no matter how high or low may be the level of the constant inflation rate. We may then say that the NAIRU is a characteristic of the real economy, whereas the inflation rate of the nominal economy is a result of past historical developments, which may involve more than the short-run price dynamics. In this example, where intervention of supply or adjustment factors are however neglected, the NAIRU appears well defined, independently of any convention.

But how far can this example be generalized? In order to answer the question, it is convenient to give to (253) a matrix form, so as to make easy comparison with the much more general system (231)–(232). Our present example is a particular case with:

$$P_p(L) = P_w(L) = W_w(L) = 1, \qquad W_p(L) = L$$

and we write it as:

$$\begin{bmatrix} 1 & -1 \\ -L & 1 \end{bmatrix} \begin{bmatrix} g_{pt} \\ g_{wt} \end{bmatrix} = -\sigma \begin{bmatrix} 0 \\ u_t - \hat{u} \end{bmatrix}. \tag{255}$$

Here we see that the square matrix in the left-hand side has not a well behaved inverse: the inverse of its determinant $(1 - L)$ has not a convergent power expansion (or, to be more exact, let us say that the power expansion derived from (255) does not give convergent solutions in g_{pt} and g_{wt} for all sensible past evolutions of u_t). So, the method we have used for the solution of (231)–(232) does not exactly apply on the example. However

$$\begin{bmatrix} 1 & 1 \\ L & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ -L & 1 \end{bmatrix} = \begin{bmatrix} 1 - L & 0 \\ 0 & 1 - L \end{bmatrix}. \tag{256}$$

So, premultiplication of both sides of (255) by the extreme left matrix of (256) leads us to

$$\begin{bmatrix} g_{pt} - g_{p,t-1} \\ g_{wt} - g_{w,t-1} \end{bmatrix} = -\sigma \begin{bmatrix} u_t - \hat{u} \\ u_t - \hat{u} \end{bmatrix}, \tag{257}$$

which indeed gives us (254).

Let us now go back to system (231)–(232), neglect for convenience the terms in z_{pt} and z_{wt}, finally linearize $\pi(d_t)$ and $\sigma(u_t)$ like in (246) so as to write:

$$\begin{bmatrix} P_p(L) & -P_w(L) \\ -W_p(L) & W_w(L) \end{bmatrix} \begin{bmatrix} g_{pt} \\ g_{wt} \end{bmatrix} = \begin{bmatrix} \pi(d_t - \hat{d}) \\ -\sigma(u_t - \hat{u}) \end{bmatrix}. \tag{258}$$

The generalization of the example concerns cases in which the square matrix of this system has no well behaved inverse because the determinant (235) is equal to zero[114]. In such cases it is possible to find a square matrix $M(L)$ so as to have an equivalent to (256), in the sense that:

$$M(L) \begin{bmatrix} P_p(L) & -P_w(L) \\ -W_p(L) & W_w(L) \end{bmatrix} = \begin{bmatrix} 1-L & 0 \\ 0 & 1-L \end{bmatrix}. \tag{259}$$

Thus, the system (258) implies:

$$\begin{bmatrix} g_{pt} - g_{p,t-1} \\ g_{wt} - g_{w,t-1} \end{bmatrix} = M(L) \begin{bmatrix} \pi(d_t - \hat{d}) \\ -\sigma(u_t - \hat{u}) \end{bmatrix}, \tag{260}$$

which generalizes (257).

The reader may see how this operation applies to the following case:

$$\begin{aligned} &P_p(L) = W_w(L) = 1, \qquad P_w(L) = \frac{1-b_p}{1-b_pL}, \\ &W_p(L) = \frac{1-b_w}{1-b_wL}, \end{aligned} \tag{261}$$

where b_p and b_w are two positive numbers smaller than 1: there is lagged but complete indexation of prices on wages and of wages on prices. The matrix $M(L)$ for this case can be read in the form that the solution system (260) then take, namely:

$$\begin{aligned} \begin{bmatrix} g_{pt} - g_{p,t-1} \\ g_{wt} - g_{w,t-1} \end{bmatrix} &= \left[b_p + b_w - b_pb_w(1+L)\right]^{-1} \\ &\times \begin{bmatrix} (1-b_wL)(1-b_pL) & (1-b_p)(1-b_wL) \\ (1-b_w)(1-b_pL) & (1-b_pL)(1-b_wL) \end{bmatrix} \begin{bmatrix} \pi(d_t - \hat{d}) \\ -\sigma(u_t - \hat{u}) \end{bmatrix}. \end{aligned} \tag{262}$$

[114] We note in passing that the econometric fits given in Gordon (1997, op. cit.) do not significantly reject a case of this type.

We note that the condition for all $g_{pt} - g_{p,t-1}$ to be equal to zero is:

$$\pi d_t - \sigma u_t = \pi \hat{d} - \sigma \hat{u}. \tag{263}$$

A permanent excess demand on the market for goods may be compensated for by a permanent excess supply on the market for labour. Similarly, changes in the NAIRU will not result only from supply or adjustment permanent shifts affecting the wage equation, i.e. those captured by the variable z_{wt}, but also from those affecting the price equation and captured by the variable z_{pt}. This would be made explicit if we reintroduced in (262) the variables z_{pt} of Equation (231) and z_{wt} of Equation (232). This shows that indeed the NAIRU is a more useful concept for the theory of inflation than for the theory of unemployment, which must focus on the labour market.

Two models are thus underlying the two interpretations of the definition of the NAIRU, those exemplified here respectively by Equation (242) and by the system (262). The distinction rests on whether the "long-run" determinant (235) is positive or nil. It fundamentally depends on what we mean by the long run in a macroeconomic theory of short-run movements in prices and wages. If the claim is to predict what would be, after several decades, the impact of demand pressures on the rate of inflation of an economy in which other factors, represented here by z_{pt} and z_{wt}, would remain constant and neutral, clearly the second interpretation is more convincing: with permanent excess demands, the rate of inflation would permanently increase and sooner or later become unsustainable. But a model of which the relevant horizon would be of a decade or less might very well have a positive determinant and be appropriate. In any case the difference in interpretation does not much matter if it is between two modelling choices which would fundamentally differ only by value of the determinant (235), it being positive but small in one case, nil in the other (c hardly smaller than 1, or $c = 1$ in the specification (243), for instance).

4. Now that we know the theoretical background given at present to the NAIRU let us look with hindsight at the historical development of ideas since four decades. The "*Phillips curve*" played an important part in it; but reference to the curve also at times obscured the debate about the determinants of inflation. To understand exactly the significance of the Phillips curve, we have to define not only what it really is, but also the conditions under which it applies.

A. Phillips observed that, over a century, there was a simple empirical link in the United Kingdom between annual changes in the rate of the nominal wage

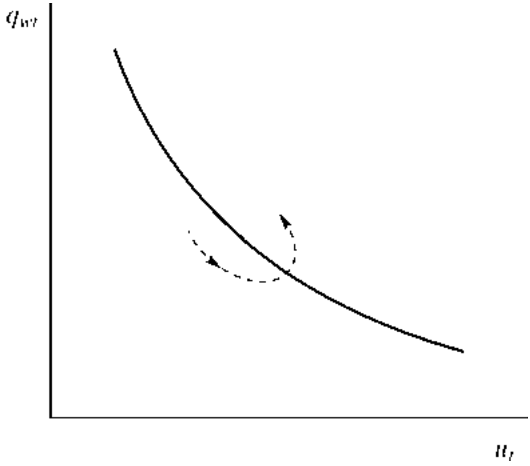

Fig. 7.

and in the rate of unemployment[115]. In the notation chosen here, a relation such as the following traces this empirical finding:

$$g_{wt} = \sigma(u_t), \tag{264}$$

where σ is a decreasing function (Figure 7). The relation was later popularized as linking the rate of inflation to the rate of unemployment; thus it was expressed as:

$$g_{pt} = \varphi(u_t), \tag{265}$$

where φ was also a decreasing function. In this form it gave an extremely simple formulation of the dilemma faced by macroeconomic policy: not only did it stipulate that, in order to reduce unemployment, government had to accept that inflation would accelerate, but it also claimed to provide a numerical rela-

[115] A.W. Phillips, The relation between unemployment and the rate of change of money wage rates in the United Kingdom, 1861–1957, *Economica* (November 1958). J. Tobin notes that actually the same empirical relationship had been earlier found by I. Fisher on American data in: A statistical relationship between unemployment and price changes, *International Labour Review* (June 1926) 785–792. See J. Tobin, Supply constraints on employment and output: NAIRU versus Natural rate, Cowles Foundation Discussion paper No. 1150 (April 1997).

tionship; this showed in particular which rate of unemployment was associated with absence of inflation (the solution of $\varphi(u_t) = 0$).

Later statistical observation undoubtedly proved that a law as simple as (265) did not universally apply; indeed, the 1970's saw a sharp rise in unemployment in many countries and at the same time a marked acceleration in inflation.

Clearly, the model exposed in the previous section explains not only why we may observe a Phillips curve in certain circumstances, but also why this is not always the case. To deduce Equations (264) and (265) from (229), (231) and (232), we just need to consider the case in which two particularities would hold. On the one hand, no exogenous change, other than those acting through pressure of demand, would affect the process of price–wage formation: not only would functions $\pi(d_t)$ and $\sigma(u_t)$ remain unchanged, but also the variables z_{pt} and z_{wt} would keep constant values. On the other hand, adjustments in prices and wages would fully take place within the period of demand shock: the functions $E(L)$, $P_p(L)$, $P_w(L)$, $W_p(L)$ and $W_w(L)$ would come down to constants.

Taking into account lags in adjustment complicates the relationships somewhat without fundamentally changing their implications, at least if the lags are short, which is not always the case, as we shall see. It was indeed realized very quickly that lags were required if econometric fits had to be at all tight, especially when attention turned from annual to quarterly series. Thus the "Phillips law" then meant that in a given quarter the rate of inflation was related to the rates of unemployment observed not only during the current quarter but also during a few earlier quarters. In the diagram with u_t on the horizontal axis and g_{pt} on the vertical axis, observed points did not trace exactly the downward Phillips curve but rather loops around it: a clockwise loop when indeed g_{pt} appeared to react with a lag to changes in u_t. Now on the contrary a counter-clockwise loop is more commonly recognized, a phenomenon which can be explained by the idea that lags in price–wage adjustments are less important than lags of unemployment with respect to pressure of demand and by the observation that the change $u_t - u_{t-1}$ in unemployment plays a moderating part in the wage equation, besides the part played by unemployment itself.

As soon as it had been published, the article of A. Phillips had stimulated economists to search for explanation of the empirical law and for contrary evidence. Particularly noticed at the time was an article of R. Lipsey, who discussed various aspects of the Phillips' article and posed various questions for further research[116]. Among those questions was the role of price expectations,

[116] R. Lipsey, The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1862–1957: a further analysis, *Economica* (February 1960).

which attracted much attention in the subsequent years. The result was the so-called "augmented Phillips curve", whose equation was:

$$g_{pt} = g^e_{pt} + \varphi(u_t), \tag{266}$$

when it was written directly for price inflation. A change in price expectations would shift the whole curve, upward in case of an upward revision of expectations. We note in passing that this equation is very close to those, like (39), which were introduced here in Section 2.2 and subsequently underlay a number of developments, particularly those in Part 3.

Introducing expectations was also a way for rationalizing the presence of lags in the Phillips law. Indeed, Equation (266) directly connects to the main model discussed here if price expectations are adaptive. When, for example, expectations obey (43), the rate of inflation implied by (266) is determined by either:

$$g_{pt} = \hat{g} + \frac{1 - bL}{1 - cL}\varphi(u_t) \tag{267}$$

if a is smaller than 1, i.e. if expectations are regressive, where c is defined by (46), or by:

$$g_{pt} - g_{p,t-1} = (1 - bL)\varphi(u_t) \tag{268}$$

if $a = 1$. The distinction made earlier here between the two interpretations of the NAIRU crops up again.

Following this line of argument, proponents of the augmented Phillips curve were also led to distinguish a "short-run Phillips curve" which linked g_{pt} to u_t, given earlier values $u_{t-\tau}$, from a "long-run Phillips curve" related to the stationary level of the rate of inflation which would correspond to a constant value of the rate of unemployment. If (267) applies, the slope of the first curve is given by φ', that of the second by $\varphi'/(1 - a)$. In so far as a is close to 1, the sensitivity of inflation to unemployment is much larger in the long run than in the short run. An expansionary policy which would reduce the rate of unemployment could very well lead, in the immediate future, to a quite moderate acceleration in inflation, but also could induce, little by little as expectations would adapt, an intolerable rate of inflation.

A landmark in the penetration of the line of argument was when M. Friedman claimed that the long-run Phillips curve was vertical and that so a persistent policy of demand expansion was self-defeating[117]. He was then accepting

[117] M. Friedman, The role of monetary policy, *American Economic Review* (March 1968).

the augmented Phillips curve with non-regressive adaptive expectations (the case when $a = 1$ here above). If $\hat{u}$ is the abscissa of the long-run vertical line, attempts at maintaining the unemployment rate at a different level from $\hat{u}$ will eventually fail because of their implications on prices and nominal wages: permanently accelerating inflation for a target u_t lower than $\hat{u}$, permanently accelerating deflation in the opposite case. This is precisely the NAIRU under the second interpretation discussed earlier here, even though we did not make it appear to be dependent on a particular hypothesis about the formation of price expectations. Friedman called it "the natural rate of unemployment", a concept toward which we must now turn our attention.

5. In the macroeconomic language, the use of the phrase "*the natural rate of unemployment*" is often found loose, to the point that some economists abstain from it (indeed, in this book, it does not often appear). If it has nevertheless a definite attraction, it is because of about the same vision as is inspiring the neoclassical synthesis: the unemployment rate would tend to gravitate towards a natural level. Thus, the concept is really an assertion about the real world. This is clear in the first introduction of the word by Friedman (1968, op. cit.), when he makes the comparison with the older distinction between the actual rate of interest and the natural rate of interest. But, once the assertion is made, one or the other of the following two questions spontaneously comes: how do you measure this natural rate? How do you explain this natural rate? Answers to the questions have often been considered as definitions of the concept, and from there confusions arose.

For believers of the assertion the best measure may be the long-run average of the actual unemployment rate. But since tendency to the natural rate is generally thought to be slow, the average ought to be computed over a fair number of decades; the measure would still be exposed to substantial errors after a series of favourable or adverse shocks, which would have long counteracted the gravitation to the natural rate. Moreover, nobody believes that the natural rate would necessarily remain unchanged after important structural or institutional changes directly or indirectly affecting the labour market; so the average computed over a long past period may have little relevance given current structures and institutions. We shall come back to the difficulty when considering in Chapter 9, Section 2.9, attempts at decomposing the actual unemployment rate into two components, the "cyclical rate" and the natural rate.

An alternative measure is available for those who identify the natural rate with the NAIRU (we shall examine in the Section 6.6 how the NAIRU can be econometrically estimated). But this amounts to assuming that equations meant to trace short-run movements in prices and wages are also good at detecting the point of attraction for the unemployment rate. The assumption, which the author of this book finds unwarranted, cannot be justified without reliance on a theory of the long run.

Here we hit the main difficulty: in the same way as the neo-classical synthesis did not thus far exhibit a unanimously agreed theory of the steady-state growth path towards which the economy would tend, similarly the fairly widespread belief in a tendency towards a natural rate of unemployment cannot exhibit a theory of the determination of this natural rate, a theory about which believers would agree.

Certainly conscious of the difficulty, Friedman, in his famous article in which actually the NAIRU was identified with the natural rate[118], tried to be explicit about his theory of the long run and then called a sort of general equilibrium theory in. He wrote: "The natural rate of unemployment is the level which would be ground out by the Walrasian system of general equilibrium equations, provided that there is imbedded in them the actual structural characteristics of the labor and commodity markets, including market imperfections, stochastic variability in demands and supplies, the cost of gathering information about job vacancies and labor availabilities, the costs of mobility, and so on". The text is far from being fully explicit, but we may read in it a definition of frictional unemployment, a definition somewhat similar to the one we tried to give in Chapter 7, Section 3.1, without claiming however that the rate of unemployment would tend to it.

Other economists, more motivated by the theory of unemployment than by that of inflation, have left the NAIRU concept aside and tried to give independent theoretical explanations of a natural rate. We shall later refer to parts of this research work. But we did already refer to other parts, when discussing in Chapter 7 a number of equilibrium theories looking at a medium rather than at a short term; recalling them will suffice for our present needs.

Some of the models then considered were intended to transpose to a longer horizon the Keynesian equilibrium usually viewed as applying for analysis of the next year or so; others did not concentrate on aggregate demand deficiency but rather on an inherent tendency of the socio-economic system to generate a too high real wage rate in the absence of excess supply of labour; this could occur because of the form of wage bargaining, of the need for employers to offer efficiency wages, or of the role of insiders; other models still showed the possible roles of imperfect competition or of coordination failures. In each one of these models, an equilibrium level of unemployment was defined in

[118] The same identification is suggested in another article written at about the same time and also speaking of a natural rate of unemployment: E. Phelps, Money-wage dynamics and labor-market equilibrium, *Journal of Political Economy* (July–August 1968).

excess of frictional unemployment. As we stressed at the beginning of Section 5.6 of Chapter 7, it was then legitimate to use the phrase "equilibrium unemployment" in order to refer to the level so defined. But the concept was then abstract, in that it was clearly model-dependent.

Unfortunately, a new source of confusion arises with the practice of some economists to now speak of "the equilibrium unemployment" as if it concerned the real world. Indeed, the phrase is then used without reference to any theory or model. Why do not these economists speak of the natural rate, which is already available for their purpose? The confusion illustrates how difficult it is for human communities to stick to a precisely defined language: over this century at least, the meaning of the word equilibrium in economics was the abstract one given in the first chapter of this book. We need the concept of equilibrium unemployment when we discuss the implications of an efficiency-wage theory, of an insider-outsider theory or of any other of the theories recalled above. Each one of them captures some features of the real world; but none of them provides a full description of this world; none of their equilibrium rates of unemployment gives the explanation of the natural rate of unemployment, even though each one should provide elements for this explanation.

6.4. Econometrics of the price equation

At this point we must turn our attention, even if briefly, to the results of econometric research which aimed at estimating the effects of factors likely to explain changes over time in prices and wages[119]. The reader will realize that the author knows only a sample of these results, but hopefully enough to give a valid introduction to the larger body of the relevant literature. Two econometric approaches have been used, aiming at knowledge of either a structural equation or a reduced equation like (239) or (240) here. In the second case the data came from macroeconomic time series, whereas in the first case they were more disaggregated and often even really microeconomic. We begin here with

[119] The scope of econometric studies about such changes in prices and wages is larger than what is needed for enlightening the concerns and models of this chapter. We already briefly referred to some studies of this type in Chapter 4, Part 2, and in Chapter 7, Section 3.4. Many more studies concern business cycle facts, which will play an important part in Chapter 9. It would not be convenient at this stage to rule out some results on the ground that they concern business cycles rather than the sizes of specific effects on inflation or the lags with which these effects occur. So, we shall make reference here already to some articles whose main interest lies in the theory of business fluctuations but which are also somewhat relevant for our present topic. We shall do so particularly with articles giving results of econometric investigations on panels of disaggregated or microeconomic data.

"the price equation", something like Equation (231) where the evolution of wages would be taken as exogenous; the next section will concern "the wage equation", (232) say, with an exogenous evolution of prices; Section 6.6 will deal with results about the reduced form equation, or with results of the simultaneous estimation of the whole price–wage system.

Simply summarized the main outcome of the econometric research on the price equation is that the evolution of prices closely mimick that of costs, to the point of suggesting a simple "cost-plus formula" for the determination of prices, that is, a constant average markup. There are, however, signs of systematic deviations from such a pattern and they seem to come from four groups of somewhat conflicting forces: a high level of demand gives opportunity for sellers to increase their prices; liquidity problems, which are particularly widespread in depressions, force firms to raise their prices and to accept, hopefully temporary, losses of competitiveness; during business fluctuations marginal costs, which are the most relevant for pricing decisions, do not evolve in proportion to average costs; price rigidities may prevent quick application of a constant markup. The bulk of the available evidence along these lines concerns producer prices of manufactured goods or retail trade prices. Price equations directly fit at the level of the whole economy and taking wage rates as exogenous do not seem to be able to disentangle the respective effects of these somewhat conflicting forces[120].

Even before we refer more precisely to some econometric results we know that determinants of price changes depend on the nature of the good or service sold. Producers of the major intermediate goods such as steel, metals, chemical products, etc., have to closely follow market conditions and have no interest in asking prices which differ significantly from those set by competitors at home or abroad[121]. At the other end of the scale, producers of specific types of equipments or services have quite a lot of freedom in making their price decisions, because their clients have many things to consider apart from prices when deciding on their purchases; only large and durable differences from what would be considered as normal prices might undermine, but then for a long time, the competitive position of these producers. Manufacturers of cars, household

[120] See, for instance, the results concerning a number of European countries and reported by J. Drèze and C. Bean, Europe's unemployment problem: introduction and synthesis, in: J. Drèze et al. (eds.), *Europe's Unemployment Problem* (MIT Press, 1991). See also J. Sargan, The consumer price equation in the postwar British economy: an exercise in equation specification, *Review of Economic Studies* (January 1980).

[121] However, such prices were also found to often remain fixed for fairly long time spans by D. Carlton in the study considered in Section 3.4 of Chapter 7. This may reflect the existence of implicit or explicit long-term contracts with customers.

machines, etc., have to announce their prices throughout a large distribution network, which means that their list prices are rather rigid; however these announced prices do not always apply in the same way to the actual transaction prices (discounts, credit facilities, part exchanges, etc.).

Likewise, we should not ignore the large body of propositions and hints coming from microeconomic theory. We discussed pricing in the two first parts of Chapter 4, as a component of short-term decisions of firms. We did not there refer to longer-term considerations, which also play a part and were examined in the theory of decisions under asymmetric information: a seller draws obvious advantages from having regular customers, but attachment of those customers requires trust in the fairness of pricing practices of the seller; a price appears fair if it is justified by a rise in cost, not if it is openly said to be permitted by a temporary boost of demand; the "customers-market theory"[122] thus gives a rationale for constant markups.

Indeed, one of the outcomes of the econometrics of price variations is to have even detected more in this respect: the cost to which a fairly constant markup is applied seems to have been, in most cases, not the short-run marginal cost, favoured in the simplest microeconomic theory, but rather a "normal cost" close to an average cost, or at least to a long-run marginal cost. This was found already in the 1970s for American and British manufacturers[123], and more recently as well as for Dutch retail traders[124].

Since the middle 1980s a number of econometric studies were performed on panels of US, more or less finely disaggregated, industry data in order to estimate variations in the short-run marginal costs and variations in the marginal markup, i.e. the excess of the price over the marginal cost. We gave two references at the end of Section 2.3 of Chapter 4. Others can be added[125], all giving estimates of short-run marginal costs which tend to be increasing with the volume of production. These studies have to face two difficulties: they must first define a method for the estimation of marginal costs and methods actually used vary; second, since they are using panels, they must simultaneously es-

[122] See Section 3.4 of Chapter 7.

[123] L. Sahling, Price behavior in US manufacturing: an empirical analysis of the speed of adjustment, *American Economic Review* (December 1977); K. Coutts, W. Godley and W. Nordhaus, *Industrial Pricing in the United Kingdom* (Cambridge University Press, 1978).

[124] B. Nooteboom, A. Kleijweg and R. Thurik, Normal costs and demand effects in price setting: a study of retailing, *European Economic Review* (April 1988).

[125] M. Bils, The cyclical behavior of marginal cost and price, *American Economic Review* (December 1987); I. Domowitz, G. Hubbard and B. Peterson, Market structure and cyclical fluctuations in US manufacturing, *Review of Economics and Statistics* (February 1988); J. Rotemberg and M. Woodford, Markups and the business cycle, *Macroeconomics Annual* (National Bureau of Economic Research, 1991), pp. 63–119.

timate cross-industry variations in competition and cyclical variations, when they have to assess the dominant features of variations in the markups. Because of these difficulties the results vary. For instance, I. Domowitz et al. (op. cit.) conclude that (marginal) markups are procyclical in non-durable goods industries and acyclical or slightly counter-cyclical in durable goods industries, whereas M. Bils (op. cit.) finds in most industries a definitely counter-cyclical evolution of markups and explains the difference with the conclusion of I. Domowitz et al. by the fact that these authors did not take account of variations in overtime premia when estimating the short-run marginal labour costs. Overall the conclusion that markups are counter-cyclical dominates and seems to be now accepted by macroeconomists.

Liquidity problems are often associated with involuntary accumulation of inventories. More generally these problems have been taken as responsible for the observed fact discussed above, namely that markups tend to be high in recessions[126]. Other, but probably less persuasive, explanations of this fact can be found. For instance, the simple model of short-run monopolistic behaviour studied in the first part of Chapter 4 implies that the equilibrium markup increases when the price-elasticity of demand decreases and we there argued in favour of a short-run demand function having a higher price elasticity at high volumes of exchange than at low ones[127]. It has also been argued that firms are more able to collide in recessions than in booms[128]. But it is not clear whether such explanations could square with the various reasons otherwise given for prices to be fairly rigid in the short run, except when increases in costs justify them.

Even though less immediately obvious and hardly significant in some of the fits, the pressure of demand also plays a part[129]. The various studies did

[126] For an old reference, see T. Courchene, Analysis of the price inventory nexus with empirical application to the Canadian manufacturing sector, *International Economic Review* (October 1969); for two recent ones, see J. Chevalier and D. Scharfstein, Liquidity constraints and the cyclical behavior of markups, *American Economic Review* (May 1995); J. Chevalier and D. Scharfstein, Capital-market imperfections and countercyclical markups: theory and evidence, *American Economic Review* (September 1996).

[127] See also J. Stiglitz, Price rigidities and market structure, *American Economic Review* (May 1984).

[128] See J. Rotemberg and M. Woodford, Oligopolistic pricing and effects of aggregated demand on economic activity, *Journal of Political Economy* (December 1993).

[129] See not only the references given in footnotes 14 and 15 above but also: G. de Menil, Aggregate price dynamics, *Review of Economics and Statistics* (May 1974); L. Maccini, The impact of demand and price expectations on the behavior of prices, *American Economic Review* (March 1978); J.-P. Pollin, Les prix industriels dans l'inflation, *Revue Economique* (Novembre 1977); J. Fayolle, Emploi et prix: un modèle de court terme construit sur des variables d'opinion, *Annales de l'INSEE* (Octobre–Décembre 1983).

not take the same definition of the measure capturing demand pressure: an indicator of the degree of capacity utilization was the most frequent; but other indicators include the amount of standing orders and measures derived from qualitative answers given at business test surveys (for instance, about the reason why output was not larger).

As a concrete example of the possible outcome of such econometric studies, we can quote here a price equation of the French quarterly model METRIC[130], an equation concerning industrial producers prices and estimated over the period 1963–1976. The actual price p_t is modelled as aiming at adapting to a desired price p_t^*, the speed of adjustment being all the higher as is the pressure of demand:

$$\log p_t/p_{t-1} = (0.55 - 0.13m_t - n_t)\log p_t^*/p_{t-1}, \tag{269}$$

where m_t is an indicator of the unutilized margin of industrial capacities, whereas n_t is a dummy for government policy interventions, especially price freezes. The unobserved desired price is a result of the same econometric estimation and is written as:

$$p_t^* = c_t(1 + 0.48e_t - 0.35e_{t-1}) + 0.41 f_t, \tag{270}$$

where c_t is a measure of the unit cost of production (which depends in particular on the labour cost and on import prices), whereas e_t is an index of French cost competitiveness with regard to export prices of foreign competitors and f_t is an index of how much credit rationing firms have to face. Thus liquidity problems lead to a higher desired price. Moreover the more competitive firms are, and are becoming, the higher the desired price.

Since the indicator m_t remains almost always smaller than 0.20, Equation (269) implies a fast adaptation of the price p_t to its desired level as long as no government intervention occurs ($n_t = 0$): half of the discrepancy between p_t^* and p_{t-1} is already corrected in quarter t. Other econometric results about price equations lead to a definitely slower transmission of cost changes.

6.5. Econometrics of the wage equation

Econometric studies aiming at explaining evolutions of nominal wage rates during the second part of this century have been very numerous (we do not

[130] *The METRIC Model* (INSEE, Paris, 1981).

speak here of studies concerning wage disparities at a given time, which also have been numerous). As a whole, they recognize the importance of four factors, or rather groups of factors: the increase in prices, the pressure of demand on the labour market, the correction of discrepancies between the actual level of real wages and what it ought to be, finally government incomes policies as well as other changes in the relative positions of partners in wage negotiations. Limits of the information content of data sources, multiplicity of the relevant factors, variations in the phenomenon through space and times explain why specifications explored by various econometricians differ and why their results do not always agree. A picture however emerges, which will be grossly depicted now.

Up to recently this econometric work was viewed as providing additions to, or revisions of, "the augmented Phillips curve" aimed at explaining short-run variations in the nominal wage rate. But during the last decade work on the wage equation payed less attention to the short-run dynamics and more attention to the setting of the real wage rate. In the last part of this section we shall consider econometric contributions inspired by this somewhat different approach.

1. For concrete reference in what follows let us begin with two examples of wage equations. The first one, borrowed again from the French METRIC model and concerning quarterly changes of the industrial wage rate, writes in our notation:

$$g_{wt} = 0.95\bar{g}_{pt} - 0.88\bar{v}_t + z_{wt}. \tag{271}$$

In this equation $\bar{g}_{pt}$ is a smoothed rate of growth of consumer prices over three quarters:

$$0.95\bar{g}_{pt} = 0.48g_{pt} + 0.32g_{p,t-1} + 0.16g_{p,t-2}. \tag{272}$$

Likewise $\bar{v}_t$ is a smoothed indicator of the disequilibrium on the labour market, v_t being the logarithm of the ratio between unemployment and a measure of the number of job vacancies:

$$0.88\bar{v}_t = 0.18v_t + 0.26(v_{t-1} + v_{t-2}) + 0.18v_{t-3}. \tag{273}$$

Finally, z_{wt} takes into account not only seasonal variations but also exceptional policy measures concerning wages in 1968. In explaining the fluctuations in g_{wt} between 1964 and 1976, the terms in $\bar{g}_{pt}$ and in $\bar{v}_t$ play roles of comparable importance.

The second example applies to annual series of average wage rates in nineteen OECD countries over the years 1952 to 1983[131]. A central equation, showing a kind of simple average of national results, reads here:

$$g_{wt} = 0.3 g_{w,t-1} + 0.7 g_{p,t-1} - 0.3 \log\left[\frac{w_t}{p_t q_t}\right] - 2u_t, \tag{274}$$

where q_t is the trend of labour productivity, as estimated by linear interpolation between successive peaks (notice the similarity between the third term in the right-hand member and the error correction (238)).

In both of these examples indexation of wages on the cost of living appears (in Equation (271), the estimated coefficient 0.95 does not significantly differ from 1, but for Equation (274) full indexation was assumed in the specification of the model). The indexation lag is short in both cases (actually shorter with the French quarterly results than with the average OECD yearly results).

In most studies use of consumer price index, like in those examples, is chosen. But, in some cases in which the wage equation is estimated in parallel with, or even simultaneously with, the price equation, an index of producer prices is used. Few studies only have introduced directly observed price expectations rather than current or past inflation[132]. The main point is, however, to note that full indexation is not universal and that results about indexation lags much vary. For instance, D. Gordon claimed that in the United States the average degree of indexation was limited to 78 per cent, or even to 55 per cent if a number of ancillary significant explanatory variables were introduced in the wage equation[133].

A good deal of the observed variation in the degree and speed of indexation seems to be due to the more or less acute public awareness of inflation. We already noted in Section 1.4 that price expectations were adapting particularly quickly where and when inflation was fast; we may think that the same phenomenon is naturally acting upon wage formations.

The most frequently used indicators of the pressure of demand on the labour market are related to unemployment, but alternative indicators were also used as (255)–(257) shows[134]; how unemployment ought to enter was also much

[131] See D. Grubb, Topics in the OECD Phillips curve, *Economic Journal* (March 1986).

[132] See however S. Turnovsky and M. Wachter, A test of the expectations hypothesis using directly observed wage and price expectations, *Review of Economics and Statistics* (February 1972).

[133] D. Gordon, The un-natural rate of unemployment: an econometric critique of the NAIRU hypothesis, *American Economic Review* (May 1988).

[134] It is here the place to quote D. Déruelle's research aimed at selecting the best variable for measuring the pressure of the demand for labour and at finding out how the roles of the different

discussed. For instance, Grubb (1986) notes that, with the national unemployment rate u_t in (258), the fit is often improved by the additional introduction of the increase $u_t - u_{t-1}$, which appears with a negative coefficient: this is equivalent to recognizing the existence of counter-clockwise loops around the Phillips curve, like in Figure 7.

It was also recognized that not all unemployed played equivalent parts in the competition for jobs and in the restraint of wage increases: they are not active to the same degree in their job search; they do not look to employers as equivalently employable. In particular, econometric studies, like that of Grubb (1986), noted either that the effect of the aggregate unemployment was not linear as assumed in (274), the effect actually increasing less than proportionally when unemployment increased, or that, when considered separately, the rate of long-lasting unemployment had a much smaller moderating effect that the rate of the other part of unemployment. In the same spirit, question were raised about whether the employment rate and the rate of participation in the labour force should not replace in the equation the unemployment rate, or whether that should not be a measure of flows into unemployment, or still whether factors acting on frictional unemployment should not appear.

Supporting the findings derived from econometric analysis of time series of the last few decades, other studies used different data bases. Here we can quote a scrupulous analysis of more than 3,000 collective agreements concluded in Canada between 1953 and 1973. The effect of the rate of growth in prices and the state of the labour market at the time when these agreements were made comes out very clearly[135].

Confirming the idea that the law linking, at the aggregate level, nominal wage increases to various characteristics of the economic situation could depend on the institutional context, study of historical data showed that wages seem now to be less sensitive to the state of the labour market than they were in the past, whilst sensitivity to prices is accentuated[136].

explanatory factors in the wage equation systematically changed depending on skills, industries and regions. He could then check that the role of the legal minimum wage was important for unskilled workers but not statistically significant for skilled workers. Pressure of the demand played a more important role in men's wages than in women's' (the coefficients were of the orders of 1.4 and 0.9, respectively). This fact is understandable when we realize that the female participation rate is very sensitive to the demand for labour: the female supply of labour is more elastic than men's, as we saw in Chapter 2. See D. Déruelle, Hausse à court terme du salaire nominal, tension du marché du travail et mouvements des prix et du SMIC, *Revue économique* (Mai 1975).

[135] W. Riddell, The empirical foundations of the Phillips curve: evidence from Canadian wage contract data, *Econometrica* (January 1979).

[136] R. Boyer, Les salaires en longue période, *Economie et statistique* (September 1978), J. Sachs, The changing cyclical behavior of wages and prices: 1890–1976, *American Economic Review*

Equations presented thus far do not contain any variable which could be considered as directly capturing changes of the "relative bargaining strength" of employers and employees in wage negotiations. This concept may not be so easy to measure. But some authors have used variables such as "the proportion of trade union members among wage earners" or the "the number of labour conflicts". The outcome is that higher wage inflation is indeed associated with lower social consensus[137].

Similarly incomes policies or other aspects of public policy deserve attention. The dummy variable z_{wt} in Equation (271) for French industrial wages did reflect a particular case of this sort; other studies, dealing with the French average wage rate, contain more in this respect, at least changes in the legal minimum wage rate. Attempts at introducing changes in taxation were also made in a number of countries, and the variable measuring them was sometimes found to play a significant part.

To end up with this last fourth group of factors, those most connected to national institutions and historical background, we may consider an econometric study[138] concerning 19 US manufacturing industries during the period 1954–1968. These industries were classified in two groups depending on whether their employees were strongly unionized or not. The following results were obtained:

– for the industries in which trade unions played a large part:

$$g_{wt} = a + 0.07 g_{pt} - 1.99 z_t - 0.18 u_t - 0.09 \dot{u}_t, \tag{275}$$

where u_t denotes the rate of unemployment and $\dot{u}_t$ its rate of change; z_t is a dummy variable taking the value 1 during the period when the government announced norms for wage increases (guideposts) and the value 0 during the other years;

– for the industries where trade unions were weak the estimate was:

$$g_{wt} = b + 1.04 g_{pt} + 0.33 z_t - 0.92 u_t - 0.51 \dot{u}_t \tag{276}$$

(a and b are constant).

(March 1980), A. O'Brien, The cyclical sensitivity of wages, *American Economic Review* (December 1985).

[137] See J. Mc Callum, Inflation and social consensus in the seventies, *Economic Journal* (December 1983).

[138] O. Ashenfelter, D. Johnson and J. Pencavel, Trade unions and the rate of change of money wages in U.S. manufacturing, *Review of Economic Studies* (January 1972).

The difference between (275) and (276) is striking: in unionized industries wage increases do not seem to have been very sensitive either to the increase in prices or to the amount of unemployment; on the contrary, government norms aimed at moderating wage increases had a noticeable effect on the results of collective agreements[139]. On the other hand, the influences of increases in prices and unemployment are very clear in industries with weak trade unions, for which government norms in no way slowed down the rate of wage increases.

2. The best way to see the entry of some recent econometric developments about the wage equation probably is to closely look at the error-correction terms, which are more and more often found in econometric results. Indeed, it may be a good place to remember that study of the econometric treatment of dynamic models with error correction was initiated by J. Sargan in 1964 precisely for application to the wage equation we are now discussing[140].

In Equation (274) the error-correction term related the real wage rate just to the trend of labour productivity. The relation is natural and was indeed confirmed by later econometric work. For instance, Drèze and Bean (op. cit.), commenting on results obtained for various European countries with wage equations including error corrections, report that "measured productivity gains seem to be passed quite rapidly into wages with short-run elasticities after a year ranging from 0.4 to 0.8 and with a long-run elasticity close to unity".

Introducing an error-correction term, also permits a more sophisticated notion of the equilibrium level of the real wage than in (274). Instead of a simple identification with the trend of labour productivity, a specification of this level may incorporate in particular the pressure of demand on the labour market, in conformity with the law found by D. Blanchflower and A. Oswald[141]. As we already mentioned in Chapter 4, Section 2.4, these two authors, working on many data sets, found that real wage rates earned by employees were negatively related to the unemployment rate in their region or industry. The resulting "wage curve" establishes a second robust empirical association between wages and unemployment, no longer between the change in nominal wage

[139] In fact, this conclusion about the results of collective agreements has been strongly qualified by R. Flanagan, Wage interdependencies in unionized labor markets, *Brookings Papers in Economic Activity*, No. 3 (1976). However, the main point of the results was reiterated from different data in D. Gordon (1988, op. cit.).

[140] J. Sargan, Wages and prices in the UK: a study in econometric methodology, in: P. Hart, G. Mills and J. Whitaker (eds.), *Econometric Analysis for Economic Planning* (Butterworths, London, 1964).

[141] See, for instance, D. Blanchflower and A. Oswald, An introduction to the wage curve, *Journal of Economic Perspectives* (Summer 1995).

rates and the level of unemployment, like in the "Phillips curve", but also between the level of real wages rates and the level of unemployment. If both of these laws are thought to hold, the unemployment rate may appear twice in the equation meant to apply to changes in the nominal wage rate, inside as well as outside the error-correction term; but time series may be too little informative for accurate simultaneous estimation of the two effects. In most of the results reported by Drèze and Bean (op. cit.) the unemployment rate appears only within the error-correction term.

The concept of an equilibrium real wage rate is usually related to one or another of the theories we discussed in Chapter 4: wage bargaining (Section 2.2), long-term contracts (Section 2.3) or wage policy of firms, particularly "efficiency wage" (Section 2.4). Each one of these theories may bring other variables into the picture. Indeed, the error-correction terms now commonly contain such additional variables, often meant to act on the "reservation wage" of employees: the ratio between the average level of unemployment benefits and the average wage, an indicator of the duration of unemployment benefits, the importance of withheld income tax or wage taxes, an indicator of business profitability, . . .

Faced with these recent developments, we may even wonder whether they do not lead macroeconometricians to overlook what was the purpose of earlier work on the wage–price spiral, namely characterization of the short-run dynamics of inflation. For instance, a number of interesting analyses and econometric results concerning the spiral are reported in the book by R. Layard, S. Nickell and R. Jackman[142]. But we may wonder whether these results are not too systematically merged into models aiming at different purposes, notably into a dominant bargaining model focused on wage-setting and price-setting. Awareness of rigidities and adjustment costs, which was acute in earlier works, may now be put too far aside. O. Blanchard and L. Katz[143] point to an interesting distinction in this respect between Western Europe and the US: fits to time series show that the error-correction term plays and important part in the European wage dynamics, but not in the US. This implies that the European NAIRU, but not the American NAIRU, depends on such factors as energy prices, interest rates and payroll taxes. This may be interpreted as being due to a greater role of unions in wage-setting or to more stringent hiring and firing regulations in Europe.

[142] R. Layard, S. Nickell and R. Jackman, *Unemployment, Macroeconomic Performance and the Labour Market* (Oxford University Press, 1991).

[143] O. Blanchard and L. Katz, Wage dynamics: reconciling theory and evidence, *American Economic Review* (May 1999).

6.6. Econometrics of the NAIRU

1. We now turn attention to more global estimations of the short-run price–wage dynamics, dealing either simultaneously with the price and the wage structural equations, or with a reduced form equation giving the rate of inflation. In the second case, determination of the reduced equation may or not also involve a connection between the two measures of demand pressure, respectively on the goods and the labour market (see Equation (229), for instance).

Simultaneous estimations of the price and the wage equations were made in many countries. For the USA a long study of O. Blanchard using monthly data and followed by a discussion may be taken as a example[144]. Assuming from the start that the system of the short-run dynamics is homogeneous in nominal variables, in the sense that the determinant (235) is equal to zero, Blanchard reaches two conclusions, namely that his indicator for pressure of demand does not seem to play a direct significant part in his price equations and that adjustment of prices to wages is as slow as adjustment of wages to prices, this fact being explained by addition of cost-repercussion lags along the input–output channels from primary factors to finished products. In other words, the cost-plus formula would apply, but with lags. The homogeneity assumption is disputed by C. Sims. R. Gordon, drawing from his own work, states that he finds a slower adjustment of prices to wages than of wages to prices, and that a significant demand-pressure effect on prices appears when a proper indicator of pressure on the goods market is introduced.

Another interesting reference is provided by G. Laroque and B. Salanié who, following R. Quandt and H. Rosen, introduced excess labour supply as the indicator of market tensions[145]. The four authors treat this indicator as an unobserved latent variable, so echoing the difficulty of the concept to which we referred in Section 3.1 of Chapter 7. But information is available about excess labour supply from two sources. First, variations in the observed unemployment rate may be assumed to reflect with some lags variations in this excess supply: an equation such as (229) may be assumed to hold, the pressure of the demand for goods d_t being there replaced by the labour excess supply es_t. Second, this excess supply may be explained, simultaneously with employment, by factors of supply of, and demand for, labour. The hypothesis according to

[144] O. Blanchard, Aggregate and individual price adjustment, *Brookings Papers on Economic Activity*, No. 1 (1987).

[145] G. Laroque and B. Salanié, Un modèle de déséquilibre de la courbe de Phillips en France et en Allemagne, *Annales d'Economie et de Statistique*, No. 44 (1996); R. Quandt and H. Rosen, *The Conflict Between Equilibrium and Disequilibrium Theories: The Case of the US Labor Market* (Upjohn Institute, 1988).

which employment is equal to the minimum of the supply of labour and the demand for labour can then be accepted for simplicity, as it was accepted for most of Part 4 of Chapter 7 (presence of random errors in the supply and demand leads to a specification somewhat akin to those in Section 4.8). More precisely Laroque and Salanié also assumed the supply of labour to be a function of the potential labour force, of the real wage rate and of a trend, and the demand for labour to be a function of lagged employment, of output and of the real cost of labour. Equations such as (231) and (232) then respectively gave the price equation and the wage equation, the market tension indicator u_t being replaced by es_t and its lagged values in (232), whereas d_t was simply removed from (231). Using appropriately elaborated econometric techniques the authors could simultaneously estimate the various equations of their model.

Let us look more particularly at the results obtained for the equivalent to Equations (229), (231) and (232) in their fit to German quarterly series for the years 1966 to 1989 (employment having been found blocked at the level of the demand for labour briefly in 1967–1968 and consistently since the end of 1974). Since there are quite a few parameters to simultaneously estimate, no wonder that accuracy was not tight (we should also remember those were the years of difficult revisions of people expectations in Western Europe). The following results are, however, sufficiently accurate and relevant to be quoted here:

$$u_t = 0.65 es_t + 0.35 es_{t-1} \tag{277}$$

(the level of the latent variable es_t is arbitrary)

$$g_{pt} = 0.3 g_{p,t-1} - 0.1 g_{wt} + 0.4 g_{w,t-1} + 0.02 z_{pt} + 0.001, \tag{278}$$

$$\begin{aligned} g_{wt} = {} & -1.3 g_{w,t-1} + 3.1 g_{pt} - 0.4 g_{p,t-1} \\ & - 0.6 es_t + 0.3 es_{t-1} + 0.02, \end{aligned} \tag{279}$$

where g_{pt} and g_{wt} are quarterly growth rates and where z_{pt} is equal to the growth rate of foreign prices.

If we isolate the two last equations from the model in which they were obtained, we note that the short-run price–wage dynamics is a little bumpy. However, the two equations define a stable evolution with characteristic roots equal to −0.68 and 0.45. The long-run determinant (235) is far from nil, being equal to 0.8, which is high although not quite exceptional among results about the price–wage dynamics. This means that the two equations do not determine a kind of NAIRU in the strict sense given to the word[146] in Section 6.3 (the "second interpretation" there). They are probably better descriptive of trends perceptible through the short-run dynamics than of the true long-run characteristics of the German economy; so let us speak of medium-run effects. The Phillips effect is clearly present on (279). In the medium run an increase of 1 per cent in the excess labour supply, i.e. also of 1 per cent in the unemployment rate, is associated with a decrease of 0.4 per cent in the annual rate of price inflation and of 1.0 per cent in the corresponding rate of wage inflation.

2. Separate indicators of demand pressure on the two markets may also be used in the estimation of a reduced price equation, which may be characterized here as obtained from elimination of the wage rate only from Equations (231) and (232). This is done, for instance, by H. Sneessens and J. Drèze[147] who use the unemployment rate as the pressure indicator on the labour market and an estimated ratio between the demand for domestic goods and output as the pressure indicator on the goods market. D. Staiger, J. Stock and M. Watson also are arguing that other demand pressures variables have power as leading indicators on inflation when it is forecast from a price equation already containing unemployment: such proposed extra variables are the capacity utilization rate in manufacturing, the standing unfilled orders at manufacturers of durable goods, the index of new orders of the National Association of Purchasing Managers[148].

A reduced equation of form (242) is estimated by P. Fortin for the years 1956 to 1984 in Canada[149]. The demand-pressure indicator then is the unemployment rate of adult men. Additional exogenous variables, like z_{pt} and z_{wt}, are simultaneously introduced in the equation: the growth rate of non-wage costs and dummies concerning price and wage controls from 1976 to 1978, and the 1983 government guidelines. Since the hypothesis of no structural change is not rejected, a NAIRU $\tilde{u}$ for adult men is computed according to the value which (242) would predict for a long-run equilibrium when other variables

[146] A long discussion of the German NAIRU, as it may be inferred from a more traditional detailed econometric treatment of the price and wage equations, can be found in W. Franz and R. Gordon, German and American wage and price dynamics: differences and common themes, *European Economic Review* (May 1993). The low estimated feedback from wages to prices, which appears here in (278) and is responsible for the high value of the long-run determinant (235), is also found by these authors.

[147] H. Sneessens and J. Drèze, A discussion of Belgian unemployment combining traditional concepts and disequilibrium econometrics, *Economica* (Supplement) (1986).

[148] D. Staiger, J. Stock and M. Watson, The NAIRU, unemployment and Monetary policy, *Journal of Economic Perspectives* (Winter 1997).

[149] P. Fortin, How "Natural" is Canada's high unemployment rate?, *European Economic Review* (January 1989).

would be fixed and maintained constant at their average value during the sample period. A constant unemployment rate of adult men would have implied a varying unemployment rate of the whole labour force, because of changes in its demographic structure and in welfare provisions, to which adult men are much less sensitive than others.

The performance of an equation of type (240) has been followed since many years by R. Gordon. Fitted on quarterly US series, now for the period 1955 to 1996, the equation contains the following exogenous variables besides the unemployment rate: the excess of productivity growth above its trend, the rates of increase in the relative price of imports, and in that of food and energy, finally dummy variables concerning the introduction and removal of price controls between 1971 and 1975 (see R. Gordon, 1997, op. cit.). The fit appears to be quite good.

3. We should keep in mind that US inflation during those past decades remained quite moderate, in comparison with what happened in many other countries; overall, its movements exhibited a high degree of inertia. Results concerning countries with a very different experience in this respect should be also considered. Turning now attention to *estimation of changes in the NAIRU*, we shall however refer again to the US, where this question was particularly studied.

In Section 6.3 we defined the NAIRU as "the unemployment rate that is consistent with steady inflation in the absence of transitory supply or price–wage shocks". We are now adding that it is not viewed as a constant parameter and may vary over time. Clearly, this NAIRU is not directly measured by any statistics. For an econometrician it has the nature of an unobserved latent variable, which plays an important part in the model explaining inflation. In order to estimate this latent variable, the econometrician will ask for a model in which the assumed role of the NAIRU is made explicit and its changes are allowed. Again R. Gordon (1997, op. cit.) offers such a model and relates it to his preferred equation for inflation.

In order to examine this model, let us, for simplicity, write (240) as:

$$g_{pt} = B(L)g_{p,t-1} + P(L)z_t + P_u(L)u_t + a. \tag{280}$$

Let u_t^* be the latent variable representing the, possibly time-varying, NAIRU. Transposing an equation like (280), Gordon proposes the following model:

$$g_{pt} = B(L)g_{p,t-1} + P(L)z_t + P_u(L)(u_t - u_t^*) + \varepsilon_t, \tag{281}$$

$$u_t^* = u_{t-1}^* + \eta_t, \tag{282}$$

where ε_t and η_t are random disturbances with zero means and standard deviations equal to σ_ε and σ_η. The idea behind (282) is that changes in the NAIRU, although allowed, should be fairly slow. Hence σ_η should be small. Thus, we should not be much bothered with the non-stationarity of the assumed random walk (an autoregressive process with high inertia would practically lead to similar estimates).

But why should u_t^* in the system (281)–(282) be called the NAIRU? This is clear when, first, z_t is subject only to transitory shocks and $z_t = 0$ corresponds to the normal value of this variable (this was assumed by Gordon for each one of his exogenous variables), and when, second, $B(1) = 1$ (which held in Gordon's estimation to a close approximation and could otherwise be imposed in the specification). Then, assuming in (281) $z_t = 0$ and $u_t = u_t^*$ for all t, we find that the rate of inflation will indeed oscillate around a constant value. However, permanent shifts in z_t would require that corrections be made in the estimation of u_t^* following from estimation of the system (281)–(282). For instance, a permanent shift of z_t from 0 to $\hat{z}$ would require in u_t^* the correction $\hat{u}$ given by $P(1)\hat{z} - P_u(1)\hat{u} = 0$. Since some supply shocks are often viewed as permanent, the remark should not be overlooked.

In order to apply the model to the estimation of u_t^*, something more should be assumed about the errors ε_t and η_t than was said above, because σ_ε and σ_η are not separately identifiable thus far (examining closely this point would lead us too far off). Gordon puts forward the idea that the evolution of u_t^* should be fairly smooth and that this should be the leading consideration, permitting a proper choice of σ_η, which has not to be so precise: a quick trial with few values will suffice to show what the proper order of magnitude is.

In his application to the American GDP deflator, Gordon concludes that, from 1955 to 1996, the NAIRU varied between 5.3 and 6.5 per cent, being particularly low in the early 1960s and high in the early 1980s. There are, however, reasons for thinking that the estimation of the time-varying NAIRU cannot be precise: not only would the estimates be already subject to non-negligible standard-errors if all the specification choices made by Gordon were perfectly fitting the underlying phenomenon, but these choices themselves add causes of uncertainty. For instance, we suggested above that perhaps not all shocks other than those concerning demand pressure should be treated as transitory. If they had all been treated as permanent (admittedly an opposite extreme), the estimated NAIRU would have reached something like 6.8 per cent between 1974 and 1980.

6.7. About the dynamics of unemployment

Let us pause and see where the main thread of our argument led to thus far. The objective is the study of inflation. But, particularly in Section 6.3, we had to look at unemployment as a close neighbour of inflation. So the reader has reasons to think that the theory of unemployment is intertwined with that of inflation. Indeed, both have to do with market disequilibria; the Phillips curve even suggests that the short-run dynamics of unemployment may be just the mirror image of the dynamics of inflation. How far does this association extend?

The focus of attention in this part of the chapter is the system ruling the joint short-run evolution of prices and wage rates. This is a partial system with two main versions, one defined by the two equations (231) and (232), therefore taking as given the two sequences of d_t and u_t, the other adding Equation (229), hence taking as given only one of these sequences. The system in question transposes the one defined earlier in Sections 2.1 to 2.3, so as to explicitly introduce two important characteristics of the labour market, the wage rate and the unemployment rate. In the same way as consideration of the formation of aggregate demand for goods in Section 2.4 led to a less partial system, we now have in the present framework to look at the formation of demands and supplies on both the goods and the labour markets. We shall so reach a sort of theory of both inflation and unemployment. This will be done in the next part of the chapter geared toward policy analysis. Before we turn to it, it may be useful to have in mind the inspirations motivating these further developments.

There is first and foremost the simple vision embodied in "the neo-classical synthesis": in the short-run macroeconomic evolutions are dominated by movements in aggregate demand. The main dynamics will then be inspired by Keynesian analysis, which was indeed already underlying most discussions of this part, as well as by the hypothesis made in Part 2 and stipulating that productive capacity was exogenously given. Justification for this vision will be presented in the next chapter when it will be shown that models of a moving Keynesian equilibrium well portray most economic reactions at the horizon of one or two years. This vision gives the strategic role to the formation of aggregate demand for goods: supplies are sufficient in the short run, the demand for labour is a direct consequence of the demand for goods, as was argued in Sections 4.2 and 4.3 of Chapter 7, as well as in Section 6.2 of this one.

However, models inspired by this vision do not provide deep theories of inflation or unemployment. On the one hand, we insisted several times since Section 6.2 on the point that values to be given to parameters such as the coefficients e_τ in (230) might much vary depending on market conditions; and those conditions cannot be taken as given, except at quite short horizons. On the other hand, as soon as we extend the horizon, we come across relevant questions addressed to theories, such as: how can we explain that an economy and not another is prone to inflation? How can we explain that an economy and not another has been long experiencing unemployment?

Clearly, the rather mechanical models of macroeconomics cannot give full answers to such questions. Political economy considerations have to enter, particularly for inflation. The many institutional features which make labour markets more complex than markets for goods have to enter also, particularly for unemployment. We easily understand that, when the horizon is so extended, the theory of inflation and the theory of unemployment part company, and not only because of the "long-run vertical Phillips curve".

Returning to macroeconomics proper, we just have to look back to Part 5 of Chapter 7 in order to realize the challenge faced by the dynamic theory of unemployment. We there discussed models in which processes of price and wage adjustment were not all sustaining a Keynesian depression, but some rather a classical depression. The main weakness of the models was to have specified, for simplicity, adjustments which did not incorporate feedback from evolutions to expectations, a feedback which in contrast was systematically present in this chapter. We shall not attempt now to consider again these models and make them more complex. We may, however, in what follows keep their analysis in mind. The dynamics of unemployment will be discussed more precisely in the next chapter, in particular in its Section 2.9.

7 More on Policies and Diagnosis

After explicit introduction of the labour market in the analytical framework, several reasons recommend that we turn attention again to policies. Although this introduction did not fundamentally damage the value of the model defined in Part 2 and used in Part 4 of the chapter for the study of aggregate demand policies, a reconsideration of the model may lead to a somewhat different specification, which could also be useful for the same purpose. The reconsideration may indeed look deeper into two aspects of the phenomena involved: the role of the wage–price spiral in the specification to be chosen for the adjustment process eventually linking the growth rate of prices to the pressure of aggregate demand, the effect on this demand of changes induced in income distribution by the interplay between price and wage rises (Sections 7.1 and 7.2).

A cost-push inflation may come from the fact that something went wrong in the domestic formation of prices and wages. Perhaps expectations were misdirected, as they often were when a speculative bubble developed or when people did not understand the economic implications of a change in their environment. Perhaps the country went through a period of social or political turmoil. In such cases governments often think of direct interventions that would affect the formation of prices and wages; macroeconomists in particular must know how to incorporate in their analysis prices and incomes policies that are either proposed or even adopted. Now is the proper place in this book for examining the issue (Section 7.3).

A good policy requires a good prior diagnosis. How to reach such a diagnosis varies from one case to another; whoever in charge has to be alert in interpreting early signs of change and to keep an open mind to ideas, even maybe so some which may seem queer at first hearing. It would be counterproductive to pretend to impose a pattern on diagnosis search; what then matters is a good understanding of the broad range of conceptual frameworks and accumulated knowledge, which is available for intelligent appraisal of the particular case.

But part of this asset is awareness of what happened in the past after various kinds of shocks to which our economic systems are commonly exposed. During the last two decades new methods of empirical time series analysis were developed, which are interesting in particular for the study of shocks and of their impacts (Section 7.4). We shall moreover see how the VAR approach, introduced in Part 5 of the chapter, may prove suggestive again for this new purpose (Section 7.5).

Once this will have been done, we shall be ready to discuss two issues for which identification of what are the most common shocks matters. We shall thus come back to the respective effectiveness of two methods for implementing monetary policy, by a control either of interest rates, like in Section 4.1, or of the quantity of money, like in Section 4.2. We shall finish with a survey and brief discussion of the stability issues which concern both the inflation and the growth processes.

7.1. Price adjustments and the formation of demand: a reconsideration

The basic ingredients, developed in Part 2 and used in Part 4, were on the one hand a specification of a dynamic process ruling evolution of the rate of inflation as function of the evolution of the pressure of demand, Equation (47) for instance, on the other hand, a specification of a law to which changes in the pressure of demand obeyed, Equation (58) in general terms. We shall take in turn each of these two ingredients. Our discussion about the first will be little conclusive and will not lead us to perceive a clear need of revision[150]. About the second, the conclusion will in contrast look definite.

1. We have to compare Equation (47), the rate of inflation being now denoted g_{pt}, with the implication of the system defined by (229), (231), (232) when g_{wt} and u_t are eliminated, which leads to an equation of form (239). To ease the comparison of the two dynamic relations between the pressure of demand and the rate of inflation, we shall omit the exogenous variables z_{pt} and z_{wt} and

[150] So the reader may decide to skip the first part of this section.

replace $\pi(d_t)$ and $\sigma(u_t)$ by their linear approximations as given by (246). We are so led to the following equation:

$$g_{pt} - \hat{g} = D(L)\big[\pi W_w(L) - \sigma P_w(L)E(L)\big]d_t, \tag{283}$$

where $\hat{g}$ gathers all constant terms and $D(L)$ is given by:

$$D(L) = \big[P_p(L)W_w(L) - P_w(L)W_p(L)\big]^{-1}. \tag{284}$$

Equation (47) should not, however, be directly compared with (283) since we have modified the definition of the pressure of demand between Parts 2 and 6, as is shown by the comparison between Equations (32) and (227). With the definition given to d_t in Section 6.2 Equations (47) and (48) imply:

$$\big[1 + (k_1 - k_2)L\big](1 - c'L)\big(g_{pt} - \hat{g}\big) = (k_3 - k_5L)(1 - b'L)d_t \tag{285}$$

(the coefficients b and c in Part 2 are written here b' and c' so as to avoid confusion with those introduced in special models such as (243), which we shall use again later). We recall moreover that in Part 2, leading now to (285), we chose a rather precise specification of adjustments in prices and the degree of capacity utilization, together with another precise specification of adaptive expectations, whereas we preferred in part 6 more general specifications, which could accommodate many cases.

Once this is realized it appears that Equations (283) and (285) are of the same type. For comparison of the lags they respectively imply we could consider many different cases. But examination of two extreme ones should suffice for our purpose.

In Section 6.2 we considered various situations, one in particular in which, because of acute pressure on productive capacity, the pressure of demand reacted strongly and quickly on prices. In the model of Section 2.2 this implies that coefficient k_1 should be taken as equal to zero (hence $k_5 = k_2k_3$). Equation (285) then leads to:

$$(1 - c'L)\big(g_{pt} - \hat{g}\big) = k_3(1 - b'L)d_t$$

that is:

$$g_{pt} - \hat{g} = k_3d_t + k_3\frac{(c' - b')d_{t-1}}{1 - c'L} \tag{286}$$

which clearly shows the adjustment lags. In situations like this, we should expect that π and still more σ will be high, but that the power series expansion of $E(L)$ will have very low coefficients. It is legitimate to ignore in (283) the term containing $\sigma E(L)$, all the more so because, as we may be assuming here, pressure on productive capacity should mainly concern plant and machinery (the coefficients in $E(L)$ being quite small) rather than labour (σ remaining moderate). Under such conditions Equation (283) leads to:

$$g_{pt} - \hat{g} = \pi D(L) W_w(L) d_t.$$

If we take the example when (243) holds, we find[151]:

$$g_{pt} - \hat{g} = \pi d_t + \frac{\pi c(1-b)^2}{(1-bL)^2 - c(1-b)^2} d_t. \tag{287}$$

It is possible for the distributed lags implied by this equation to be similar to those which result from (286). Closer examination does not reveal existence of a systematic simple bias distinguishing those produced from one equation from those produced by the other.

On the other hand, in order to consider a situation of depression, we can focus on the case where $k_1 = k_2 = 1$, hence $k_5 = -(k_4 - k_3) < 0$. Equation (285) should then be replaced by a similar equation to (286) except that the right hand term should be multiplied by $1 - k_5 L/k_3$: the adjustment lags are a little longer than in (286), but by no longer than one period (the effect of the same change in the pressure of demand is obviously much less than in the previous case). In Equation (283) the induced change in unemployment now plays an important part. Since the values of π and σ are now lower, we can choose as a simple example, the case where $\pi = 0$ and, in addition to (243):

$$E(L) = \frac{-h}{1-bL}. \tag{288}$$

151 Here and in what follows when using specification (243), we shall limit attention to the case where wages are completely indexed to prices, although with lags, but where the increase in wages does not engender, even in the long run, a proportional price increase ($c < 1$). This implies that the effect of any temporary shock on the rate of inflation will tend to fade out. In this respect, the system's behaviour will be similar to that of those cases in Part 2 where expectations were regressive ($c' < 1 - b'$). It would obviously be possible to also treat as in Section 6.3 the case when $c = 1$; the behaviour would then be similar to that obtained in Part 2 for the case where expectations were purely adaptive ($c' = 1 - b'$): the rate of inflation would be significantly and permanently affected by any temporary shock that had happened in the past.

Equation (283) then takes the form:

$$g_{pt} - \hat{g} = \frac{\sigma c(1-b)h}{(1-bL)^2 - c(1-b)^2} d_t \tag{289}$$

which also includes a slightly longer lag than that which resulted from the equation for the previous case.

To sum up, this rapid examination does not suggest that taking explicit account of the price–wage spiral should lead to a radical change in the adjustment process which was set out in Part 2 for the reactions of the inflation rate to changes in the pressure of the demand for goods.

2. In Part 2 we gave a fairly detailed analysis of the different components in the formation of the aggregate demand for goods. Reinterpreted within the framework of this section, the resulting equation can be written:

$$d_t = A_r(L)r_t + A_p(L)g_{pt} + A_t. \tag{290}$$

A comparison with Equation (58) should take into account two changes: the definition of the pressure of demand d_t has been revised; the degree q_t of capacity utilization no longer explicitly figures in the model. But, since q_t was linked to d_t, we can easily obtain an Equation (290) which is appropriate for the new definition of d_t derived from equations set out in Part 2.

The lags implied by the specifications of the new functions $A_r(L)$ and $A_p(L)$ depend on all the factors examined so far. But the main justifications for including the interest rate r_t and the inflation rate g_{pt} in Equation (287) are the same as with Equation (58): on the one hand, that investment depends on the expected real interest rate, and on the other hand, that an excess of actual inflation over expected inflation has a depressive effect on consumption.

To these two impacts on the formation of demand we should now add the redistributive effects implied by the fact that prices and wages may not move exactly in pace. So we shall replace (290) by:

$$d_t = A_s(L)(g_{wt} - g_{pt}) + A_r(L)r_t + A_p(L)g_{pt} + A_t. \tag{291}$$

A particularly rapid increase in wages, compared with the increase in prices, has a favourable effect on the purchasing power of households' disposable income, and hence on their consumption. The consumption function (59) does not take this into account, so we have to introduce a corrective term. Yet this increase has an unfavourable effect on firms' profits, hence on their financial resources, hence on the level of their investment. Equations (60) and (62) specifying the determination of investment do not take this into account, hence the need for a second corrective term.

If these two terms exactly cancelled each other out, the changes in income distribution would not affect aggregate demand. But it is commonly said[152] that, in the short run at least, the first corrective term dominates the second, since the disposable income of households is spent more quickly than the operating income of firms. In fact we are going to assume that the coefficients of the power series expansion of $A_s(L)$ are all positive or zero, and this on the basis of the following argument.

If s_t denotes the share of wages in the aggregate income of period t, the contribution of redistributive effects to the pressure of demand can be written:

$$C(L)s_t + I(L)(1 - s_t) - 1, \tag{292}$$

where $C(L)$ and $I(L)$ characterize the lags involved in the use of income for consumption and investment, respectively. Since income from labour is spent more quickly, the coefficients of the first terms of the series expansion of $C(L)$ are higher than the corresponding coefficients of the expansion of $I(L)$. But since all incomes are eventually spent, we can assume here as a first approximation[153], that $C(1) = I(1) = 1$; this implies that $(1 - L)$ can be factored out of $C(L) - I(L)$. Consequently, (292) can be written:

$$\frac{C(L) - I(L)}{1 - L}(s_t - s_{t-1}). \tag{293}$$

Apart from changes in quantities, which are otherwise captured in (291), the change $s_t - s_{t-1}$ is approximately equal to $g_{wt} - g_{pt}$; thus $A_s(L)$ must itself be equal to $[C(L) - I(L)]/(1 - L)$, a function whose power series expansion must have mainly positive coefficients.

As an example, we can choose:

$$C(L) = \frac{1 - b_c}{1 - b_c L}, \qquad I(L) = \frac{1 - b_i}{1 - b_i L}, \tag{294}$$

[152] As we saw in Chapter 2, Section 2.8, the empirical proofs of the proposition are not as strong as many economists believe.

[153] The discussion will be resumed in Chapter 9, Section 1.5, with slightly different hypotheses, directed more to the study of the effects resulting from cyclical changes in the labour share s_t.

where the numbers b_c are b_i are such that: $0 < b_c < b_i < 1$. The result is:

$$A_s(L) = \frac{b_i - b_c}{(1 - b_c L)(1 - b_i L)} \tag{295}$$

whose series expansion indeed includes only positive coefficients.

7.2. Aggregate demand management

How do effects of changes in income distribution affect the conclusions previously obtained regarding aggregate demand policy? To answer this question we have to compare the dynamic properties of the process which we obtained in the foregoing section with those of the process we would have obtained if we had assumed $A_s(L)$ to be zero.

Everything relies on how $g_{wt} - g_{pt}$ the rate of change in the real wage, depends on d_t, the pressure of the demand for goods. This dependence is the result of the wage–price spiral, here formalized by Equations (231) and (232). Aside from the constants and the exogenous variables z_{wt} and z_{pt} which will play no role in the present discussion, these two equations imply:

$$g_{wt} - g_{pt} = D(L)\big\{\big[P_p(L) - P_w(L)\big]\sigma(u_t) - \big[W_w(L) - W_p(L)\big]\pi(d_t)\big\}. \tag{296}$$

Once u_t is expressed in terms of d_t through (229), this equation gives a good definition of the dependence in question[154]. Nevertheless, it does not provide any immediate answer to the question as to whether $g_{wt} - g_{pt}$ is an increasing or decreasing function of d_t.

Moreover, we should recognize here the drawbacks of the simplification which led, in this part as well as in the preceding one, to eliminating the distinction between the pressure of the demand for goods and the utilization of capacity. A very high pressure can be maintained either because equipments are working near full capacity or because the labour supply is almost completely used. These two situations lead to different implications of Equation (296). In order to know these implications we have to take the particular situation we are faced with into account. To fix our ideas, let us consider the following four cases.

[154] The major weakness of (296) is to ignore the presence of an error-correction component such as (238) in the wage equation. In the long run this component would certainly play a significant role. But we do not pretend here to go beyond the short run.

(i) If the degree of utilization of capacities is low and the rate of unemployment is low, any increase in the pressure of demand for goods clearly reflects on the demand for labour and, even though the unemployment rate cannot change very much, $\sigma(u_t)$ has such a high elasticity that we observe an acceleration in the growth of real wages when the pressure of demand increases.

(ii) We should still expect a positive reaction when unemployment is high if the degree of utilization of equipments is low, since changes in the demand for goods clearly affect unemployment. The sensitivity of $\sigma(u_t)$ is certainly lower than in the previous case but dominates that of $\pi(d_t)$, according to econometric studies.

(iii) On the other hand, we expect a negative reaction if unemployment is high and the degree of utilization of capacities also high. In this case changes in the pressure of demand for goods have little impact on the demand for labour and will affect mainly the speed of price inflation.

(iv) The only case in which there is no clear conclusion is when a high utilization of capacities coincides with a low unemployment. If pressure of demand is high price inflation will obviously accelerate; but any repercussion on the demand for labour also provokes an acceleration of wage inflation.

Anyway, if we stick to a linear approximation of the relation in question, we can write:

$$g_{wt} - g_{pt} = S(L)d_t + \hat{g}_s, \tag{297}$$

where it is understood that $S(L)$ will lead to mainly positive coefficients if capacities are relatively little used in the situation studied, but negative coefficients in case of high level of utilization and high unemployment.

Just as Equation (297) expresses the dependence of the real wage with regard to the pressure on demand, so the following equation summarizes the relationships examined in the preceding section to express the dependence of the rate of price inflation:

$$g_{pt} = V(L)d_t + \hat{g}. \tag{298}$$

Without real restriction to the significance of the analysis we assume that the series expansion of $V(L)$ has only positive coefficients. Hence, Equations (291), (297) and (298) determine how a policy acting on the interest rate affects the rate of inflation. Assume, for example, that A_t is constant, these equations then lead to:

$$g_{pt} = A_r(L)V(L)\left[1 - A_s(L)S(L) - A_p(L)V(L)\right]^{-1} r_t + a. \tag{299}$$

The main issue here is to know whether the effects resulting from changes in the distribution of incomes play an actually significant part in this relationship, for they should then be explicitly taken into account in the study of aggregate demand policies: we indeed ignored this possible source of complication in the discussions of Part 4. The most important task is then to examine the role of the term $A_s(L)S(L)$. We may limit attention to cases of stability, after the discussions we earlier had on the consequences and likelihood of this property: we so assume that a change in the interest rate does not trigger a divergent process[155].

By way of a short conclusion we say that the action of the interest rate on inflation is magnified when positive coefficients predominate in the series expansion of $S(L)$, that is, when we consider case (i) or (ii) discussed above[156]. This is easy to understand. An increase in the interest rate slows down the demand for investment; in the cases in question this slows down the increase in real wages; the resulting change in the distribution of incomes depresses aggregate demand, which reinforces the impact otherwise attributed to the increase in the interest rate. On the other hand, in the case of high utilization of capacities and high unemployment (case (iii) above) the initial slowing down of demand has a favourable effect on real wages and is then somewhat counterbalanced by the induced increase in consumption.

7.3. Prices and incomes policies

Since the main characteristic of inflation is a seemingly indefinite escalation of prices and wages, people naturally think that, if government would directly

[155] The assumption is necessary to give meaning to Equation (299), more precisely to the inverse of the term in square brackets. It assumes in particular, that the sensitivity of wages to the state of the labour market is not too high. We shall come back to the possible difficulty in the following section.

[156] The stability condition implies in particular:

$$X(1) = 1 - A_s(1)S(1) - A_p(1)V(1) > 0.$$

The long-run effect of r_t on g_{pt} is stronger when $X(1)$ is weaker, hence when $A_s(1)$ is stronger if $S(1)$ is positive (or $A_s(1)$ weaker if $S(1)$ is negative).

intervene on some prices and wages, it could slow down the process. Such is the idea motivating a "prices policy" or an "incomes policy". That a policy of this type is feasible comes from the fact that the inflationary spiral is not absolutely rigid, as we saw in Section 1.1. Except during war episodes, it is seldom possible to abruptly break the inflationary spiral, but it is often possible to slow down its course.

This sort of intervention raises many questions, which for the most part are beyond the scope of macroeconomics; so much so that many economists in the past have refused to recognize prices and incomes policies as having any value. But many cases were observed in which generalization, acceleration or persistence of fast increases in prices long continued, even in depression times; most often direct interventions on prices or wages were part of the policy package that eventually stopped rapid inflation. Given this observation it is difficult to justify the extreme position, which accepts no other way of fighting inflation than through demand management alone, or even just through monetary policy. Yet even if we do not hold this position, we should not be led into underestimating the many difficulties of prices and incomes policies[157].

1. Effective control of prices, which aims at reducing their rate of increase, has hypothetically to interfere with the spontaneous formation of these prices. Given the information that administrative services responsible for the controls in fact have, this interference is fairly blind; it usually introduces distortions which harm the efficient allocation of resources; it fosters the private inclination to escape from controlled sectors and to engage in economic activities that are difficult to control or have only secondary usefulness. Statutory reductions in wage increases may bring with them similar consequences.

Moreover, a policy acting systematically on either prices or wages interferes with income distribution. In a moment we shall consider the consequences of this for inflation. But these are obviously not the only consequences of importance, since distorting income distribution can work against certain long-term, or short-term, objectives: better real wages for labour, or the opposite requirement of maintaining profit margins with a concern for competitiveness and growth.

Moreover, what are the legal and administrative means allowing direct control of prices or wages by government? In many Western countries they are quite limited. A policy aiming at so achieving a general slow-down in the increases in prices and wages thus has often to use persuasion rather than coer-

[157] There is a literature on prices and incomes policies. It will not be surveyed here. The interested reader may for instance refer to H. Rockoff, *Drastic Measures: a History of Wage and Price Controls in the US* (Cambridge University Press, 1984).

cion: guideposts are announced on what should result from private pricing decisions; or government take part in wage negotiations between labour unions and employer associations. In many places, such actions can be effective only in exceptional circumstances, when the situation engenders a special sense of civic duty.

In more normal times prices and incomes policies act through ordinary public instruments: prices of public services, taxes and subsidies. But blocking public prices, lowering tax rates or making a public contribution to covering the costs of some goods or services obviously creates a net loss of income in the government budget; deciding to do so can be envisaged only temporarily or, in favourable cases, by anticipating measures that would have been taken later because of other considerations. Government must, therefore, wonder whether the time is right for such measures and whether public opinion is ripe for accompanying them by restraint in private pricing and wage bargaining.

These measures, most of which not only act directly on prices and wages but also change the balance of public finance, should be understood as having two facets. They interfere with the increases in prices, but at the same time also with the formation of aggregate demand. The two facets must be taken into account when we want to evaluate the usefulness of these measures in the fight against inflation. Incidentally, some commonly used measures meant to act on aggregate demand may also have a direct impact on price formation: a restrictive monetary policy pushes interest rates up, which increases costs and thus constitutes a factor pushing to price increases (remember that in Section 5.6 we found some evidence of an initial positive response of the price level to a positive shock on the interest rate).

As to prices and incomes policies, the problem that macroeconomics has to study is precisely that of the induced effects that direct action on prices or wages can provoke through aggregate demand. We ignored these in Section 7.1 above when discussing the formation of aggregate demand. We must now wonder whether conclusions there reached are still adequate for judging the final effect of controls on prices and wages. Depending upon the measures used and depending upon the nature of market tensions in the economic situation considered, the induced effects may vary greatly in direction and importance. This is exactly what we have to understand. In order to prepare examination of simple models for the purpose, let us first consider a few examples.

A general control of prices that succeeds in slowing down price increases but does not affect wages brings about a shift in income distribution in favour of real wages and thus an increase in aggregate demand. This more or less thwarts the slowing down of inflation. If the pressure of demand is already high, it is even possible that the final result may be an acceleration in inflation.

An action which successfully limits wage increases without affecting price formation affects aggregate demand in the opposite direction. The anti-inflationary action is reinforced because of this, especially in a situation of economic boom.

A decrease in indirect tax rates without any offsetting measure increases the public deficit. It brings about an even higher surge in aggregate demand than would a control of prices that would have had to same direct effect on z_{pt}. In this case, indeed, real incomes increase without the counterpart that decrease in profits would oppose to the increase in real wages; the increase in households consumption then plays its full role without being in any way offset by a drop in firms' demand.

2. In order to better understand the role of such induced effects, we shall refer to the model set in Sections 7.1 and 7.2, but we shall simplify it and eliminate everything which is not essential to the question we are discussing now.

We shall assume then:

$$A_r(L) = A_p(L) = A_t = 0. \tag{300}$$

We shall also assume that the measure under consideration directly acts only on prices, that is through z_{pt}, and we shall thus choose $z_{wt} = 0$. To simplify the formulas we shall moreover limit attention to specifications (243), (246) and (288). We shall even assume that the pressure of demand affects only the labour market:

$$\pi(d_t) = 0. \tag{301}$$

Writing (291) after these simplifications, we obtain:

$$H(L)d_t = -b(1 - bL)A_s(L)(1 - L)z_{pt} + a, \tag{302}$$

where a is an appropriate constant and:

$$H(L) = (1 - bL)^2 - c(1 - b)^2 - \sigma h\big[(1 - c)(1 - b) + b(1 - L)\big]A_s(L). \tag{303}$$

If the sequence of values of the exogenous variables z_{pt} is given, we can calculate by means of Equation (302) the sequence of values of d_t. The sequence of values of g_{pt} is then given by the equation corresponding to (289) when z_{pt} is no longer taken as zero, namely:

$$g_{pt} - \hat{g} = \frac{(1-bL)^2 z_{pt} + \sigma c(1-b)hd_t}{(1-bL)^2 - c(1-b)^2}. \tag{304}$$

The role of the induced effects on aggregate demand appears in terms where σ features.

Observe, first of all, that the system (302)–(303) gives the result that, if z_{pt} remains constant, any stationary solution includes the same pressure of demand, no matter what the value of z_{pt}. In fact, d_t only depends on the sequence of the z_{pt} through the differences $z_{pt} - z_{p,t-1}$. In other words, if the system is stable, the induced effects disappear in the long run. This is because of two assumptions which, although used in this model, are not suitable for all cases. On the one hand, Equation (243) adopted for $W_w(L)$ assumes complete indexation, even though with a lag, of wages on prices. On the other hand, aggregate demand is assumed to depend only on real wages; now, a lasting fall in the tax rate, which would not be offset and would introduce a lasting government deficit, would stimulate firms' demand for a long time, since it would produce a lasting improvement in profits. We shall come back to the consequences of this last remark.

Even with the system written as it is for the moment, the induced effects are not negligible since they play a part in the short run and they may even make the time path unstable whilst it would not have been so without them. The condition for stability here is that the roots of the equation in x:

$$H(x) = 0 \tag{305}$$

all[158] have modulus greater than 1 (see on this subject, for example, the indications given in Section 4.1 on this chapter). This condition is fulfilled when there is no induced effect ($\sigma = 0$). But it is easy to see that it can be violated if there are strong induced effects.

In the special case of specification (295) the equation leads us to:

$$\begin{aligned}(1-b_c x)(1-b_i x)\big[(1-bx)^2 - c(1-b)^2\big] \\ -\omega\big[(1-c)(1-b) + b(1-x)\big] = 0,\end{aligned} \tag{306}$$

[158] Taking into account the form of Equation (302), the root $x = 1$ does not necessarily lead to instability.

where:

$$\omega = \sigma h(b_i - b_c). \tag{307}$$

Studying the roots of this equation shows that the stability condition is satisfied if:

$$\omega < \varpi = (1 - b_c)(1 - b_i)(1 - b). \tag{308}$$

On the other hand, if $\omega > \varpi$, then apart from exceptional cases, an initial shock on z_{pt} triggers a cumulative process on d_t and g_{pt} which is kept up indefinitely. If they are inflationary as in the case of a price control with no simultaneous interference on wages, the induced effects can finally dominate, so that the control can eventually result in accelerated inflation.

We can see from relationships (307) and (308) that the stability condition is less likely to be satisfied:

– the slower the adjustments of wages to prices and of prices to wages (b lower), that is if those rather direct effects of the prices and incomes policy are slower to show up;
– the more changes in income distribution affect the demand for goods ($b_i - b_c$ higher);
– the stronger the repercussion of changes in the demand for goods on the demand for labour (h higher);
– the more marked the sensitivity of nominal wages to unemployment (σ higher). (In a model where (301) would not be assumed, we should still take into account the sensitivity of prices to the pressure of demand for goods.)

3. To complete this discussion let us still consider the case of a fall in tax rates, which not offset would bring about an increase in private incomes (of firms and households) and not just a change in the distribution of these incomes. So, two effects are combined: on the one hand, the transitory decrease in z_{pt}, which produces, in particular, repercussions through the distribution of incomes, on the other hand, a permanent exogenous increase in private incomes, hence in demand.

In the model chosen here where, in particular, the effect of a transitory shock on the rate of inflation tends to fade away in the long run[159], the first effect is

[159] This is obviously not the only possible case, as we saw in Section 6.3 when discussing the second interpretation of the definition of the NAIRU. Anyone who wants to go deeper into the subject can examine what would happen in the case of total indexation of prices on wages (and

only temporary. As the second effect is permanent, it will certainly dominate in the long run. What we need to know now is whether it will do this late and weakly or soon and strongly.

In order to give a simple qualitative examination of this question, let us consider the case in which a temporary fall in z_{pt}, say $-\varepsilon$ in period 1, and a permanent increase in A_t from period 1 on $(+\alpha)$ would occur starting from a stationary situation. The changes δd_t and δg_{pt} would be given by (302)–(304) which would then take the following forms:

$$\begin{aligned} H(L)\delta d_t &= b(1-bL)A_s(L)(1-L)\delta z_{pt} \\ &\quad + \left[(1-bL)^2 - c(1-b)^2\right]\alpha, \end{aligned} \tag{309}$$

$$\delta g_{pt} = \frac{(1-bL)^2\delta z_{pt} + \sigma c(1-b)h\delta d_t}{(1-bL)^2 - c(1-b)^2}, \tag{310}$$

where δz_{pt} would be zero for all periods except for $t=1$, the time at which it would be equal to $-\varepsilon$.

Considering the case in which the stability condition would hold, we easily determine the formula for the long-run perverse effect:

$$\delta g_{pt} \to \frac{\sigma c(1-b)h\alpha}{H(1)}. \tag{311}$$

In particular, if $A_s(L)$ is given by (295), the limit is written:

$$\delta g_{pt} \to \frac{\sigma c(1-b_c)(1-b_i)h\alpha}{(1-c)(\varpi - \omega)}. \tag{312}$$

It is obviously an increasing function of σ.

7.4. The empirical analysis of macroeconomic time series

The concern for a better empirical knowledge of shocks to which economic activity is exposed developed during the last two decades; it motivated a number of research projects; it undisputably enriched macroeconomic knowledge. We saw that it now inspires part of consumption and investment theories, which involve stochastic properties of the flows of household incomes (Sections 1.8

of wages on prices) even with a lag. In particular, he or she could change the specification chosen here by assuming $c = 1$ in (243), a case which we have explicitly excluded.

to 1.10 of Chapter 2) and of firm sales (Sections 1.3 and 3.5 to 3.7 of Chapter 4). It also appears in new developments intended to contribute to a better analysis of the full economic system.

At the global level we saw from the beginning of Chapter 7 (Section 2.1) and of this chapter (Section 2.1) the importance of a proper specification of the adjustments of prices and quantities to changes in the environment. In those two chapters we had many opportunities to realize that spontaneous reactions of the economic system and proper interventions of economic policy much depended on what were precisely the unexpected changes and the configuration of market disequilibria with which they impacted. We also had many opportunities to realize that, on such matters, empirical evidence from macroeconomic data was wanted because transposition of microeconomic knowledge was ambiguous or had to be supplemented by a proper diagnosis about the nature of the unexpected changes.

Whether for the rigorous study of behaviour of a representative agent, or for that of reactions of the global economy, or still for that of economic policy confronted with uncertainties, stochastic models had to be more and more frequently introduced. During a first phase of this theoretical development simple standard assumptions were made about the dynamic properties of the random terms or errors appearing in these models: errors were formalized as directly ruled by either pure white noise processes or, more generally, by autoregressive stationary processes. When introduction of long-run movements was deemed necessary, deterministic trends were supposed to suffice. But adequacy of such standard assumptions had later to be questioned, which led, first, to empirical research on better representations of observed time series, second, to renewal of earlier theoretical models, third, to revision of some econometric procedures.

We must now look at the empirical side of these recent developments. In conformity with what was done since the beginning of this chapter, we shall limit attention here to cases where just very few variables have to enter the analysis, fuller enquiries being left to the next chapter. Before moving to the possible role of vector autoregressions for identification of shocks, we must consider other important issues in the empirical approach to global macroeconomics.

1. In the early days of empirical macroeconomics it was common to *focus on the decomposition of time series* into trends, cycles and seasonal fluctuations, while also recognizing that some erratic disturbances remained. Such a pattern seemed to be suited for the purpose at the time. Heuristic arguments were most often accepted as providing a rationale for the decomposition method used. Depending on the case the trend or the cyclical component was the object

of the main interest. One or the other was then analysed within deterministic theories, which were not always even precisely formalized.

Fifty years ago econometricians were arguing for the introduction of stochastic specifications, which were then perceived as required for a rigorous study of the methodology of time series analysis. In particular, a simple linear model served to clarify the issues concerning extraction of the trend of a series, whether interest concerned the trend itself or the "detrended" series. For a variable x_t observed in successive periods t the trend was supposed to be linear, so that attention was brought to an equation such as:

$$x_t = \alpha t + \beta + \varepsilon_t, \tag{313}$$

α and β being two numbers to be estimated and ε_t a residual unobserved component. The trend was then $\alpha t + \beta$. Some flexibility of course remained in the use of this equation, through the exact choice of the variable x_t; for instance, it was not meant to be real output but rather the logarithm of this output, or even the log of per capita output.

For a full specification we moreover had to say something about the unobserved series $\{\varepsilon_t\}$. It was naturally meant to be the realization of a stochastic process, which had an unconditional zero mathematical expectation. Although more elaborate hypotheses were made in some applications, a standard case assumed the process to be stationary, and even to have a linear autoregressive representation of the form:

$$\varepsilon_t = \sum_{\tau=1}^{\infty} \gamma_\tau \varepsilon_{t-\tau} + \eta_t, \tag{314}$$

η_t being an unobserved white noise process and the γ_τ being numbers to be estimated.

The specification defined by (313)–(314) was particularly convenient for forecasts of the series $\{x_t\}$. Indeed, assuming the series to have been long enough for an accurate estimation of α, β and the γ_τ, as well as for a computation of close approximations to past values of $\varepsilon_{t-\tau}$, one could in period T also recursively compute good forecasts of the future values of the ε_t and x_t, according to the following formulas:

$$x^e_{T+\theta} = \alpha^*(T+\theta) + \beta^* + \varepsilon^e_{T+\theta}, \tag{315}$$

$$\varepsilon^e_{T+\theta} = \sum_{\tau=1}^{\theta-1} \gamma^*_\tau \varepsilon^e_{T+\theta-\tau} + \sum_{\tau=0}^{\infty} \gamma^*_{\theta+\tau} \varepsilon^*_{T-\tau}, \tag{316}$$

where α^*, β^*, γ_τ^*, $\varepsilon_{T-\tau}^*$ are the estimated values of the parameters and disturbances, whereas $x_{T+\theta}^e$ and $\varepsilon_{T+\tau}^e$ are forecast values (the second sum on the right-hand side of (316) will actually be truncated at such high value of τ as would make $T-\tau$ fall outside the observed periods).

A close look at the properties of the forecasts so obtained, however, may lead us to wonder whether the specification (313)–(314) is not actually still more restrictive than the assumed linearity already shows. Indeed, stationarity of the $\{\varepsilon_t\}$ process implies that the result $\varepsilon_{T+\theta}^e$ of recurrence (316) tends to zero as θ increases indefinitely, when common methods of estimation are applied. In other words the trend $\alpha t+\beta$, being deterministic, would give us reliable forecasts at long horizons.

Thinking over this feature macroeconomists realized that the secular movements of important variables need not be well modelled by deterministic trends. Such long-run movements may also be subject to shocks whose impacts do not dissipate with the passage of time. As a simple example let us substitute for (313) the somewhat similar:

$$x_t - x_{t-1} = \alpha + \varepsilon_t, \tag{317}$$

where ε_t is again a stationary autoregressive process with form (314). Then the good long-run forecasts will naturally be such that, as θ increases, $x_{T+\theta}^e - x_{T+\theta-1}^e$ tends to α^* and $x_{T+\theta}^e$ tends to be equal to $\alpha^*\theta$ plus a number that will be a function depending mostly on the last few observed values x_T, x_{T-1}, $x_{T-2}, \ldots$ Since these observed values do not usually lead to the number corresponding to the fitted trend, i.e. $\alpha^* T+\beta^*$, the forecast will be clearly different from that obtained with a deterministic trend. Moreover the successive long-run forecasts of $x_{T+\theta}$, namely the ones made in T, $T+1$, $T+2, \ldots$, will randomly fluctuate in a way reflecting the fluctuations of x_t. Users of such forecasts will have to be well aware of a source of uncertainty which would not occur with a deterministic trend.

Once the possibility of alternative specifications had been fully realized, the question was to know whether some macroeconomic time series would not be better represented by such "difference-stationary" stochastic processes as the one implied by (317) for x_t, rather than by the more traditional "trend-stationary" forms of type (313). C. Nelson and C. Plosser precisely looked at the question, using for the purpose tests introduced a little earlier by mathematical statisticians[160]. Dealing with long historical annual time series of fourteen US macroeconomic indicators, they found just one case, the unemployment

[160] C. Nelson and C. Plosser, Trends and random walks in macroeconomic time series: some evidence and implications, *Journal of Monetary Economics* (September 1982).

rate for the years 1890–1970, in which a difference-stationary representation of non-stationarity was rejected by a test against the trend-stationary representation[161]. They recognized, however, that their test had no power against the alternative of high first-order autocorrelation of the disturbances ε_t around a deterministic trend.

2. *The contribution of Nelson and Plosser was an important landmark in the evolution of ideas and methods about the empirical analysis of macroeconomic time series.* Two subsequent developments are worth considering. The first one concerned the question of knowing whether the test was robust. Many macroeconomists observed that they would not use constant linear trends in analyses involving long historical periods; moreover they would often approximately know when and how the trend changed; naturally the test was less conclusive when the alternative was, for instance, a piece-wise linear deterministic trend[162]. Some econometricians exhibited cases in which difference-stationarity could clearly be rejected[163]. Many macroeconomists concluded that, for the purpose of applications on which they were working, about the same conclusions would follow whether they would use difference-stationary or trend-stationary representations, as long as the latter would contain highly serially correlated disturbances.

The second important development was the reconsideration of appropriate econometric procedures for macroeconomic time series analyses. The favourable large-sample properties attributed to commonly used procedures depended on the realization of some conditions assumed in econometric theory, and those conditions often ruled out difference-stationarity, more generally they often ruled out that a characteristic equation associated with the random processes would have a unit root (for instance, Equation (237) in λ is the characteristic equation of the system (231)–(232) if it is perturbed by just a two-dimensional white noise; if $\lambda = 1$ is a solution, g_{pt} and g_{wt} are difference-stationary, except for the role of the exogenous variables of the system; the common large-sample theory of econometric procedures then no longer ap-

[161] In the case of the unemployment rate, the preferred alternative turns out to be pure stationarity of x_t rather than trend-stationarity.

[162] See, for instance, P. Perron, The great crash, the oil price shock and the unit root hypothesis, *Econometrica* (November 1989); J. Campbell and P. Perron, Pitfalls and opportunities: what macroeconomists should know about unit roots, in: O. Blanchard and S. Fisher (eds.), *NBER Macroeconomic Annual 1991* (Cambridge University Press, 1991).

[163] F. Diebold and A. Senhadji, The uncertain unity root in real GNP: a comment, *American Economic Review* (December 1996).

plies[164]). In this book we shall not enter into a serious study of the reconsideration of econometric methodology and of its implications. We shall just give a warning in a moment, but before that we shall also point to important generalizations of the simple univariate difference-stationary process introduced by (317)–(314).

Three types of generalizations are suggested by earlier developments of this chapter, particularly in the modellizations discussed in Part 6. To realize it, we just note that we then considered specifications applying to, for instance, the rate of growth in prices g_{pt}; but the focus could be on other related variables. If x_t would in (317) designate not g_{pt} but instead the logarithm of the price level p_t, a stochastic specification such as (317) would mean that (to a close approximation) $\{g_{pt}\}$ was a stationary process, since $\log p_t$ would then be difference-stationary. The importance given to equations such as (242) suggests a first kind of generalizations, namely one in which the unconditional mean of the process would be more complex than $\alpha t + \beta$ in (313) or α in (317); it often involves exogenous variables such as z_{pt} or u_t. A second kind of generalization recognizes the possibility of higher order differencing than in (299); indeed, corresponding to, for instance (254), we shall often want to consider processes in which the random part of g_{pt} will be (first-order) difference-stationary; that will mean that the random component of $\log p_t$ will be "second-order difference-stationary". The process of g_{pt} is then said to be "integrated of order 1", and the process of $\log p_t$ is then "integrated of order 2".

The third kind of generalization consists in transposing from the "univariate" modelling of the evolution of one macroeconomic variable to multivariate models concerning the simultaneous evolutions of two or several variables. Such multivariate models, concerning mainly g_{pt} and g_{wt}, were the main theme of Part 6, in which we discussed at length the possible forms and implications of the system (231)–(232), including cases leading to a system such as (260) written on the differences $g_{pt} - g_{p,t-1}$ and $g_{wt} - g_{w,t-1}$. Corresponding to those deterministic models, as many stochastic specifications can be defined, which will generalize those introduced above.

Multidimensionality opens new possibilities for stochastic dependence. In particular two random processes, such as those ruling $\log p_t$ and $\log w_t$ may be said "cointegrated" if they are both integrated, but if a linear function involving them is either stationary, or at least integrated of a lower order. This will naturally occur in stochastic formalizations of models in which an error-correction component appears. When in Section 6.2 we considered the introduction of

[164] We are here touching questions concerning not only econometric theory but also the theory of stochastic processes, which may have to be well understood when economic models themselves are specified. The point will hopefully appear in Section 2.1 of Chapter 9.

component (238) in the wage equation, we noted that its effect would indeed be to peg the long-run evolution of the real wage rate to a fixed exogenous pattern; the stochastic equivalent here is that the process of $\log w_t - \log p_t$ is trend-stationary, not difference-stationary although the processes of $\log w_t$ and $\log p_t$ are both difference-stationary. Similar forms of cointegration, but involving other exogenous variables than just time, correspond to different error-correction components which may appear in the wage equation, as we saw in Section 6.5.

These developments in macroeconomic time-series analysis contributed to motivate the search for new tools of statistical inference. It was relevant to know how to test whether the stochastic process supposed to generate an observed time-series had a unit root, or how to test whether the two-dimensional process underlying the simultaneous evolutions of two time-series was cointegrated, or still how to revise common significance-tests in time-series regression analysis when unit roots, perhaps joined with cointegration, was suspected. It was natural to find these new tools presented in recent books on time-series analysis and included in software econometric packages.

It was also natural to see some macroeconomists using the new tools in dealing with long time-series. But, for a decade or so, we could also see abuses, revealing that the true stakes had been misunderstood: when a test did not reject existence of a unit root, some implied that common tests of significance had become uninformative; others seemed to claim that all the relevant information of a sample had been retrieved after a test showing that cointegration could not be rejected and after an estimation of a supposedly cointegrating vector; still others dismissed as invalid any time-series econometric work in which unit-root and cointegration tests did not appear, no matter how short was the period covered and how low would therefore have been the power of the tests in question.

3. Another problem for empirical macroeconomists was to know how to correctly take advantage of the new conceptual stochastic framework. It was not immediately realized that *identifiability could not always be taken for granted*. Whereas the decomposition of a series into a deterministic trend and the realization of a stochastic component was conceptually clear cut, the distinction could be blurred if the trend itself was to be seen as realization of a stochastic process. In multivariate analysis it was tempting to think about a multiplicity of random shock processes, each one with a natural economic interpretation; but how could these individual meaningful processes be identified?

Moving toward considering such issues, let us begin with the *distinction between the permanent and transitory components* of time series. The distinction naturally comes to mind of macroeconomists, who used it in several contexts.

Does the distinction naturally emerge from what was said earlier in this section? Not really. If the relevant stochastic process is trend-stationary, like with (313), we are tempted to say that the permanent component is the trend $\alpha t + \beta$, the stationary disturbance ε_t being the transitory component, no matter how highly serially correlated it may be according to (314). But if the process is difference-stationary, like with (317), we are tempted to say that the series has no transitory component, even if ε_t is a pure white noise, or worse if it exhibits high negative serial correlation. Somehow, this does not seem to be the correct solution to the problem posed.

Indeed, when economists speak of permanent and transitory components they usually have in mind some simple stochastic model of the following type:

$$x_t = \xi_t + \eta_t \tag{318}$$

in which the random processes $\{\xi_t\}$ and $\{\eta_t\}$ are independent, η_t being, for instance, a white noise or being specified in such a way as to have low serial correlation, ξ_t being a deterministic trend or a random walk with drift, for instance. The specification then gives an identifying restriction, which permits decomposition of x_t into its permanent (ξ_t) and transitory (η_t) components. But there is in general no reason to assume that the variable of interest will meet the particular specification chosen. The macroeconomist therefore lacks a reference for the decomposition he or she had entertained to make[165].

Perhaps in the future macroeconomists will agree on using precise conventions for decomposing any time series into a "permanent" and a "transitory" components, or for measuring the "degree of persistence" of the erratic fluctuations of any time series. The choice of such conventions may then appear to enrich the tools of empirical macroeconomics. But, for the time being, let us conclude that the decompositions proposed have to be defined within stochastic models that are explicitly assumed to hold.

7.5. To which shocks are economies commonly exposed? Answers from the VARs

1. Let us now turn to the *identification of various types of shocks in multivariate analysis* and to the role that Vector-Auto-Regression techniques can then play. The stakes are clear: throughout our discussions in Chapter 7 and this chapter, we recurrently made distinctions between demand shocks, supply shocks and

[165] For a fuller discussion see D. Quah, The relative importance of permanent and transitory components: identification and some theoretical bounds, *Econometrica* (January 1992).

adjustment shocks (which most often are shocks to the process of formation of prices and wages). As a background for our macroeconomic diagnoses, we should like to know which kinds of shock occurred in the past. How could we use the empirical VAR approach for the purpose?

When considering the VAR technique in Sections 5.5 to 5.7 we saw how it related the series of observed variables to those of unobserved disturbances, innovations or shocks supposed to have determined actual evolutions. We also saw how the relation could be characterized by a set of impulse-response functions. We moreover realized the difficulty of identification of the system and why lack of identifiability would jeopardize the relevance of most results coming from the approach. This is why we discussed the identifying restrictions used in the search for an empirical test of monetary policy effectiveness.

The same difficulty of identification concerns applications of the empirical VAR approach to the question posed in the heading of this section. We shall consider here two alternative sets of identifying restrictions which were used for the purpose. The prior ideas embodied in the first set of restrictions concern the interpretation to be given to the contemporaneous correlations between the variables analysed. We shall not discuss the role of these ideas at present, while looking at the results obtained by O. Blanchard and M. Watson[166], because we shall do so in Section 4.4 of Chapter 9 when we shall more generally examine the approach of what are now called the "structural VARs". For the time being let us consider the suggestive results without questioning their robustness.

The authors were addressing two questions. The first concerned the number of sources of shocks: was there only one dominant source or were there many, each appearing as having played a major part at some times? The question is obviously relevant, for instance with respect to "monetarism", which may be identified with the vision that monetary shocks are the main source of economic fluctuations. The second question concerned the way the shocks were leading to large fluctuations: were fluctuations caused by an accumulation of small shocks, each being unimportant if viewed in isolation, or were fluctuations due to infrequent large shocks? The first alternative inspires a research line initiated by Ragnar Frisch (see Section 2.1 of Chapter 9); the second alternative underlies many descriptions and policy discussions in which particular economic fluctuations are ascribed to particular large shocks followed by periods in which the economy would return to equilibrium, except that another large shock may later occur.

In order to answer the two questions the authors had quarterly US data for the years 1947 to 1982 and looked at the simultaneous movements of four

[166] O. Blanchard and M. Watson, Are business cycles all alike?, in: R. Gordon (ed.), *The American Business Cycle* (University of Chicago Press, 1986).

macroeconomic variables: the logarithms of real GNP, of the GNP deflator and of nominal M_1, together with an index of the fiscal policy effect on aggregate demand. They focused attention on four structural residuals obtained from a VAR, these being interpreted as measuring respectively the aggregate demand shocks, the aggregate supply shocks, the monetary policy shocks and the fiscal policy shocks (nominal money and the fiscal index are viewed as determined by two "policy rules" taking account of output and the price level, but perturbed by respective policy shocks). Taken as an estimate of the model generating the movement of the four variables, the vector autoregression may be used to characterize the contribution of each type of shock to the unexpected movement of each variable (the unexpected movement is the difference between the actual value of the variable and the forecast that would have been given by the model on the basis of observations available, say four or twenty quarters in advance).

This VAR-analysis led to the conclusion that all four types of shock played a role in economic fluctuations, thus a conclusion opposed to the idea of one dominant source of shock. For instance, 54 (respectively 37) percent of the variance of unexpected output four (respectively twenty) quarters ahead was due to aggregate demand shocks; but the contribution of supply shocks was not negligible (16 and 20 per cent at the two horizons), neither that of fiscal policy shocks (15 and 27 per cent), nor that of monetary policy shocks (16 and 17 per cent). In the variance of the unexpected movement of the price level twenty quarters ahead the contributions were respectively: 26% for demand shocks, 22% for supply shocks, 15% for fiscal shocks, 36% for monetary shocks.

Inspection of the four time series of structural residuals led to conclude against the pure small-shock hypothesis: evidence of the presence of large shocks was found because too many residuals had an absolute value exceptionally high when compared to the standard deviation of the series to which they belonged. For instance, in the series of demand structural residuals, 4 out of 140 were larger in absolute value than 2.5 times the empirical standard deviation of the full series of demand residuals, whereas a Gaussian distribution of the shocks, to be expected in case of pure small shocks, would have led to only 2 such outlying residuals. The numbers of similarly defined outliers in the three other series were found to be 5, 5 and 6. Over the four series of residuals simultaneously, 13 were found to exceed three times in absolute value their respective standard deviation, whereas only 1 such large outlier could be expected if the distribution of shocks had been Gaussian. On the other hand, there were enough small shocks to also rule out a theory implying that only infrequent large shocks would occur. We rather see an economy exposed to a mixture of large and small shocks.

Looking more closely at the four time series, Blanchard and Watson noted that they provided some support to commonly held interpretations of the US business cycles, some support but not complete support (looking at the precise data reproduced in the article, the reader may even find that the authors in their assessments are a bit too generous about this degree of support). There was evidence in 1951 and 1952 of a sharp increase in private demand and of fiscal shocks associated with the Korean War. There was some evidence in favour of the usual explanation relating the major fluctuations of the 1970s to the two oil shocks: the importance of the supply and fiscal shocks in the middle of the decade was noticeable. Similarly, large monetary policy shocks in the years 1980 to 1982 well fitted with the role commonly ascribed to a then active monetary policy. On the other hand, from 1955 to the early 1970s, large shocks were few and not easily interpretable: no large shock signaled fiscal policy or private spending corresponding to the Kennedy tax cut or the Vietman war.

2. Another system of identifying restrictions, involving long-run effects, was proposed by several authors in the late 1980s, particularly by O. Blanchard and D. Quah[167]. The objective was again the search for a characterization of the shocks commonly affecting the course of macroeconomic evolution. The authors derived conclusions from a VAR study of two US series concerning the growth rate of output (real GNP) and the male unemployment rate (aged 20 and over). The method they had introduced looked attractive to other macroeconomists, who used it on series of other macroeconomic variables.

The approach is faced with fundamentally the same difficulty as is affecting other types of application of the VAR technique, such as the one discussed in Section 5.6: results may not look so persuasive because, inspired by a narrow conception of the information available to macroeconomists, VAR econometricians work with just very few pieces of this information, adding assumptions which may fit well into their models but may appear disputable. However, even though not quite persuasive, the results are suggestive enough to stimulate reflection and be remembered, with due caution.

In this spirit, we shall here precisely look at the work of Blanchard and Quah. Later we shall more quickly consider two other applications of the approach in order to get an idea of its potentialities for future empirical macroeconomic research.

The identifying restrictions used by Blanchard and Quah are best seen by reference to the two representations (219) and (220) given in Section 5.5 of the multidimensional process $\{y_t\}$, respectively the autoregressive and the corre-

[167] O. Blanchard and D. Quah, The dynamic effects of aggregate demand and supply disturbances, *American Economic Review* (September 1989).

sponding moving-average representation, the second one defining the impulse-response coefficients $c_{ji\tau}$. In Section 5.5 we made great use of the recursive identification, which involved first the assumption that the innovation vector ε_t had the identity matrix as covariance matrix and, second, the assumption that the contemporaneous matrix of the autoregressive form A_0 had a triangular form; with two variables only, like in the present case, the second assumption meant just $a_{120} = 0$. Keeping the first assumption $E(\varepsilon_t \varepsilon_t') = I$, Blanchard and Quah replace the second by an assumption on the long-run properties of the sequence of the matrices C_τ, more precisely by:

$$\sum_{\tau=0}^{\infty} c_{11\tau} = 0. \tag{319}$$

In other words, the innovation ε_{1t} of the first component of ε_t would elicit over time responses $c_{11\tau}\varepsilon_{1t}$ of $y_{1,t+\tau}$ which would eventually balance each other.

Remember that here y_{1t} is meant to be a measure of the rate of growth of real GNP. Actually Blanchard and Quah use for the purpose the proxy given by the variation of the logarithm of real GNP with respect to the preceding quarter. If g_t is this variation and the number y_t is real output (beware here of confusion of notation with the vector y_t), then $\log y_{t+\theta}$ is equal to:

$$\log y_t + \sum_{\tau=0}^{\theta-1} g_{t+\tau}.$$

If all the $\varepsilon_{1,t+\tau}$ and $\varepsilon_{2,t+\tau}$ are nil except for ε_{1t}, $\log y_{t+\theta}$ is, because of (220), also equal to:

$$\log y_t + \sum_{\tau=0}^{\theta-1} c_{11\tau} \cdot \varepsilon_{1t}.$$

Restriction (319) therefore means that the impact of shock ε_{1t} on future output will decrease to zero as the horizon $t + \theta$ is pushed farther into the future. In other words, ε_{1t} will have no long-run effect on output. Blanchard and Quah argue that, because of this property, ε_{1t} is "the demand shock" ε_{dt} and that its orthogonal complement ε_{2t} is "the supply shock" ε_{st}, an identification which will be discussed later.

Given (319) and $E(\varepsilon_t \varepsilon_t') = I$, the matrices A_τ and C_τ of (219) and (220) are identified and can be estimated. The four sequences of the components $c_{ji\tau}$ for

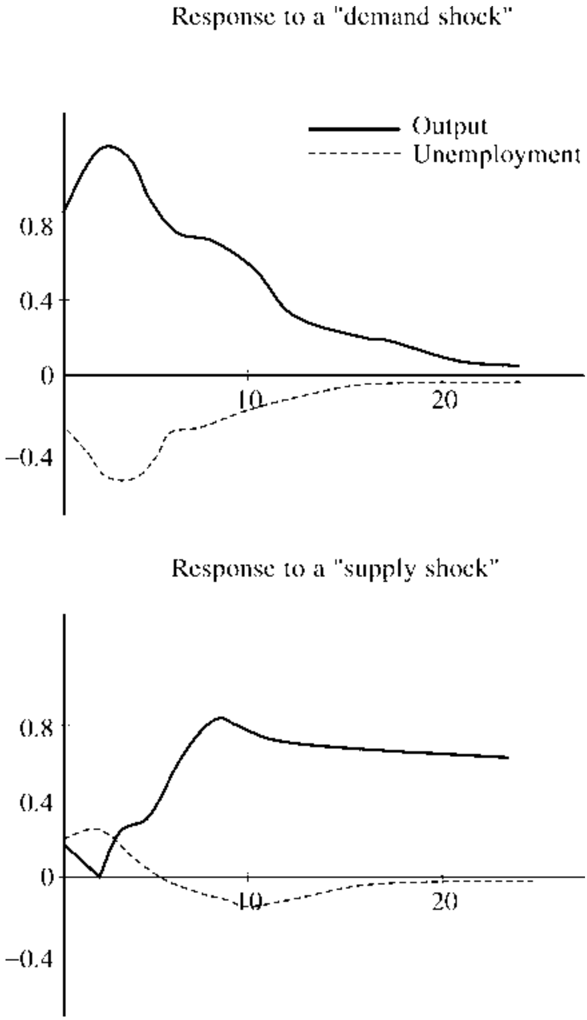

Fig. 8.

increasing τ give the impulse response functions. Figure 8 presents the main results obtained by Blanchard and Quah from the series of the 160 quarters covering the years 1948 to 1987, the autoregressive form being constrained to involve 8 lags only. (The figure concerns the results obtained from detrended series, with the assumption of a change in output growth occurring in 1973–1974; three other groups of results were also computed.)

Since we are here in the case where the autoregressive process (219) is assumed stationary and where detrended series are used, convergence to zero of the response of the unemployment rate (as well as convergence to zero of the output growth rate) has no meaning: it was supposed and not tested. The results have therefore nothing to say about a possible hysteresis of unemployment.

Blanchard and Quah note that the two impulse-response curves to their demand shocks closely fit the Okun's law (with a coefficient slightly greater than 2) but that nothing of the sort appears in the curves tracing responses to their supply shocks. The maximum output response to the supply shock occurs after two years and the long-run output response remains not far below the maximum.

Interest in the results should still more concern the identification of the nature of the shocks. The authors pay particular attention to "demand shocks" and to the output fluctuations they are estimated to have induced. They observe that those shocks played a large part in depressing output after the Korean war and the two oil-shocks, and in stimulating output during most of the 1960s and from 1983 to 1987. They point to the fact that the peaks and throughs of their demand component in output match closely those identified by the traditional descriptive methodology of the NBER (National Bureau of Economic Research).

Looking at the series of the two estimated shocks ε^*_{dt} and ε^*_{st} is interesting. Figure 9 gives the results for the periods following the two main oil-price rises. Blanchard and Quah read these estimations as follows: "The recession of 1974–75 is explained by an initial string of negative supply disturbances, and then of negative demand disturbances. Similarly, the 1979–80 recession is first dominated by a large negative supply disturbance in the second quarter of 1979, and then a large negative demand disturbance a year later".

3. Can we accept as fully persuasive this empirical detection of the fundamental nature of aggregate shocks in successive quarters? Some doubts may come from reservations about the VAR methodology in general. In the particular application where one of the identifying constraints concerns a long-run response, we may be wary of the suitability of an analysis which derives all results from an initial autoregression involving lags covering just eight quarters. We may also wonder whether estimation of orthogonalized residuals, here the so-called supply shocks, is not sensitive to the choice of the period analysed and of the initial detrending of the series (the article reports a somewhat disturbing sensitivity of the estimated part of the forecast-error variance of output or unemployment due to supply shocks).

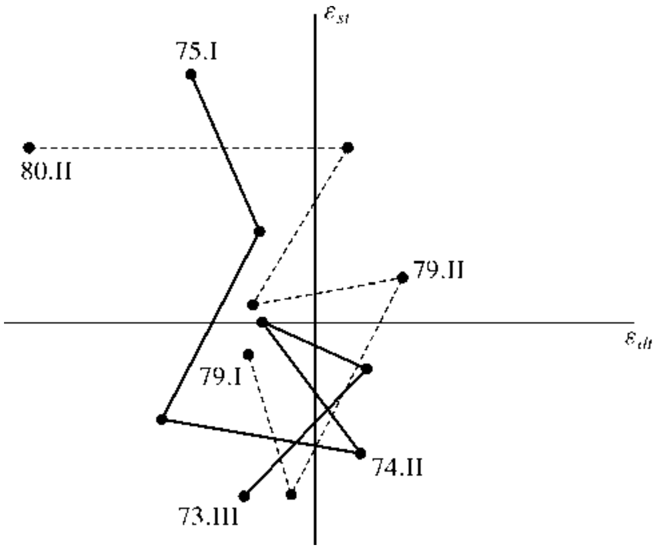

Fig. 9. Estimated shocks after oil-price rises.

The macroeconomic presumptions deserve attention. The justification of Equation (319) as implying identity between ε_{1t} and “the demand shock” comes from confidence in the neoclassical synthesis asserting that demand shocks have no long-run effects on output. Any reservation about the assertion endangers confidence in the identity, at least if the reservation goes beyond saying that there might be some such effects, while admitting that they would be small. Similarly, belief in the actual existence of a multiplicity of shocks leads to doubts about the relevance of the simple categorization according to which there is in each period just one demand shock and one supply shock: how about adjustment shocks, particularly those concerning price and wage formation? How about the idea that the latter shocks might have no long-run effect on output? In their appendix the authors provide an interesting discussion on what might happen with a multiplicity of shocks; the reader may want to refer to that discussion, which however could clearly not cover all possible cases.

The author of this book is still puzzled by the assertion of Blanchard and Quah concerning the assumption that the two shocks are uncorrelated at all leads and lags. They write that the assumption is “a nonissue” (p. 659). How can an assumption which is at the root of the whole procedure be a nonissue?

Would anyone interpreting their results as really concerning the demand and supply shocks be too naive?

We argued in Section 5.7 that VAR analyses were bringing complements of, rather than substitutes for, results brought by earlier empirical approaches. The above queries lead us to reiterate the conclusion: the results are certainly suggestive and worth remembering, but the process through which they are derived raises too many issues to be fully persuasive.

In this spirit it is worth briefly considering how impacts of the two main oil-price rises were analysed before the assistance of VAR econometrics. Two main effects were immediately recognized: terms of trade were deteriorating in net oil-importing countries; the world aggregate demand for goods and services would decline because importing countries would have to cut their demand, whereas the main exporting countries would make such large trade surpluses that they would not be able to spend them in goods and services before several years. In our present language a demand shock was clearly expected and no serious macroeconomist was surprised when the decline in world output growth was observed. Within each oil-importing country the terms-of-trade shock could be assimilated to a supply shock because it meant that the productivity of domestic labour in terms of domestic uses had declined (the resulting cost-push inflation could also be analysed in terms of a price–wage shock). The typically macroeconomist argument is still worth today; the fact that it is not contradicted by Figure 9 does not really change the degree of confidence we have in it. The fact rather acts as a test of the quality of the VAR identification of demand and supply shocks; the test is actually not demonstrative enough to lead us to forego traditional methods of macroeconomic diagnosis.

4. Let us look at two other applications of the same methodology. An extension of the Blanchard–Quah analysis was offered by E. Gamber and F. Joutz who added a third variable: the time-variation of the logarithm of the real wage rate (average hourly earnings in US manufacturing deflated by the GNP price)[168]. Given the assumption that the three innovations (or shocks), ε_{1t}, ε_{2t}, ε_{3t}, were uncorrelated, the authors needed three additional identifying restrictions, which they chose to concern the long-run responses.

Since ε_{1t} was meant to be the shock on aggregate demand (of goods and services), the neoclassical synthesis provided two restrictions: aggregate demand shocks should have no long-run effect on the real economy, which implied no such effect on the levels of both output and the real wage rate over their trends.

[168] E. Gamber and F. Joutz, The dynamic effects of aggregate demand and supply disturbances: comment, *American Economic Review* (December 1993).

Since ε_{2t} was meant to be the shock on labour supply, the authors called upon the Solow well-known 1956 growth model for the last identifying restriction. Accordingly they took it as established facts that, over the long run, the US saving rate and returns to scale had been constant, and technical progress had been neutral, except perhaps in the three cases by simple trends; hence labour supply shocks should be found to have had no long-run effect on the level of the real wage over its trend.

With these identifying restrictions they re-estimated and extended the impulse-response functions of Blanchard and Quah. They found that the responses to aggregate demand shocks closely reproduced those of Figure 8, adding a positive wage-rate response which was however proportionally less important and less peaked than the output response. Responses of output and unemployment to labour-supply shocks were roughly similar to those graphed in Figure 8 for supply shocks. During the first year or so a labour-supply shock elicited a negative wage-rate response, subsequently the response did not significantly differ from zero.

The authors also estimated responses to the third kind of shock ε_{3t}, which they called "labour-demand shock". Curiously they did not explain why they used this label or what it exactly meant as a component uncorrelated with aggregate demand shocks. A positive shock ε_{3t} was found to raise permanently the real wage and to raise, in the short run, the unemployment rate (with an unsignificant short-run decline in output). The initial assumptions of the analysis implied that the unemployment rate had no long-run response; a positive long-run response of output was found. On the basis of these results we may think that ε_{3t} captures both wage–price shocks (because of the short-run responses) and part of productivity shocks (because of the long-run responses of the levels of output and the real wage).

The authors did not present results about the time-profiles of the shocks. But they gave forecast-error variance decompositions; like in the Blanchard–Quah work, these decompositions were found to be disturbingly sensitive to the initial detrending of the series, with however a small part being systematically given to labour supply shocks.

M. Bergman[169] presents results of bivariate analyses of output growth and changes in the inflation rate for five economies (Germany, Japan, Sweden, the UK and the US). In this work, which does not involve any labour market variable and should therefore be viewed in relation with earlier parts of this chapter, the identifying restrictions besides orthogonality of the shocks bear on the

[169] M. Bergman, International evidence on the sources of macroeconomic fluctuations, *European Economic Review* (June 1996).

long-run responses: the demand shock is assumed to have no long-run effect on either the level of output or the rate of inflation, the supply shock no long-run effect on the rate of inflation.

Analysing quarterly data over a period of thirty years or so, the author fits for each country a VAR starting from an autoregression with 4 to 6 lags. The impulse-response functions to a demand shock are fairly similar, output and inflation reacting positively in the short run, with however the response of inflation being particularly small in Germany and that of output not significant in Japan and Sweden. In four of the countries the response of output to a supply shock seems to be building up during an initial period of some three years. The short-run response of inflation to the same shock is negative everywhere but insignificant in Japan and Sweden. The author points to the part played by "supply shocks", which at a five year horizon explain the larger fraction of output variation in all five countries, and also the larger fraction of changes in the inflation rate, except in Japan and Sweden. (How much of this result is due to the introduction of more restrictions on responses to demand shocks than to supply shocks is difficult to know.)

The author computes the inter-country bilateral correlation coefficients between the time-profiles of the estimated output responses. For responses to supply shocks the correlations are weak or insignificant (with an exception at 0.5 between the UK and Germany). For responses to demand shocks in contrast all but one of the coefficients are significant; the overall average coefficient amounting to 0.43. Some readers may think that correlations between estimated shocks themselves might be more instructive than between estimated output responses.

Ending this section about the nature of the shocks hitting the macroeconomy, let us note that the various authors to whom we referred assumed that shocks could be classified into a small number of categories, four at most with the article of Blanchard and Watson. Such categorizations may give a caricature of actual diversity. This was the conclusion drawn by R. Fair[170]: important shocks may come from inventories, exports, demand for housing, import prices, autonomous shifts in price formation, and so on. Moreover which shocks play the major parts depends on the length of the horizon up to which the impacts are considered.

[170] R. Fair, Sources of economic fluctuations in the United States, *Quarterly Journal of Economics* (May 1998).

7.6. The relative importance of real and monetary shocks for the choice of a monetary policy rule

1. Knowing to which shocks the economy is commonly exposed matters for policy. Indeed, fiscal and monetary interventions are best seen as belonging to strategies chosen in order to control the stochastic macroeconomic evolution following from partly unexpected changes in private behaviour and in the environment. The best strategy then depends as much on the stochastic properties of random shocks as on the structure of the systematic part of the model describing the economy. This will be a major theme in the next chapter. But we are now ready for an introduction within the simple framework of the IS-LM analysis.

This introduction will closely follow the article by W. Poole[171], which initiated a series of contributions to the same general theme. It will also provide an answer to a question that was left pending in Chapter 7, namely how to choose the instrument of monetary policy: should it be the quantity of money or the interest rate?

It has been recognized for a long time that monetary authorities have to make this choice; their interventions which feed or absorb banks' liquid assets have repercussions on the interest rate, as well as how liberal banks can be towards their clients, hence also on the liquidity of the latter, that is on the money supply. When they intervene, the monetary authorities can either keep a close watch on the interest rates to be maintained at the desired levels, or they can keep a watch on the quantity of money so that it changes at a rate chosen beforehand. If the first rule is followed, the growth of the money supply can deviate from what was expected. If the second rule is chosen, it implies corrections upwards or downwards as each monetary statistic becomes available; large and erratic movements in the interest rate can then be observed. Be that as it may, monetary authorities have to opt for one rule or the other, or even a third; they cannot choose two rules at a time.

We have realized since Section 1.2 of Chapter 7 that, in our models, we could think in different ways of the instrument of monetary policy and we did distinguish at various points between acting directly on the interest rate and acting on the money supply. We even met the idea that, from the point of view of the formation of demand which we are going to consider now, acting on the interest rate led to effects comparable to those of a budgetary measure when the interest rate is kept fixed. For example, in Part 2 of Chapter 7 we treated both simultaneously in the theory of the multiplier (Section 2.2).

[171] W. Poole, Optimal choice of monetary policy instruments in a simple stochastic macromodel, *Quarterly Journal of Economics* (May 1970).

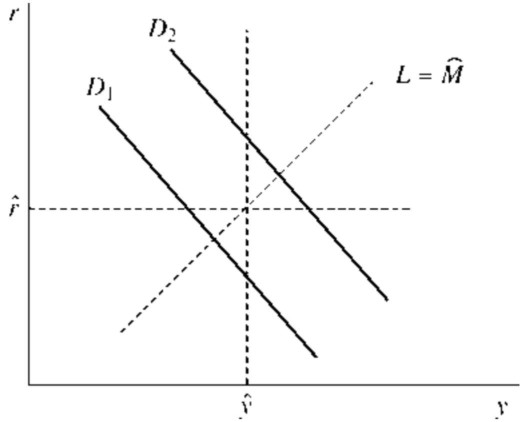

Fig. 10.

The analysis then relied on a number of simplifications, which we may find to be disputable in this chapter dealing with inflation, particularly the hypothesis that adjustments to exogenous changes bore just on quantities and not on prices, and the hypothesis that the same variable, "the rate of interest", was appropriate for discussing equilibrium in the money market as well as in the market for goods and services. We shall accept the same simplifications at present; we may now be more aware of their limitations; but we also realize that, as a first approximation, they do not seem to bias the argument in favour of one monetary policy instrument rather than of another.

More precisely we now assume that only macroeconomic effects on aggregate demand matter, and that the real interest rate can be identified with the nominal interest rate; but we take account of random shocks, in order to have another look at the theory which was the main subject of Part 2 of Chapter 7. We start from the well known diagram featuring the two curves representing the equalities $I = S$ and $L = M$ (Figure 3 of Chapter 7). On the horizontal axis, $\hat{y}$ represents the optimal level of output and aggregate demand.

2. Consider first, two simple cases. *If shocks can affect only the demand for goods and services* by households and firms, and not their demand for money, we can represent simply the possible impacts of those shocks by assuming that the *IS* curve can be either in position D_1 or in position D_2 (see Figure 10). Monetary policy has to be designed so as to achieve on average approximately the result $\hat{y}$ taking these two possibilities into account. The best choice of the

value to give to the instrument may be said to be found by referring to a curve which would lie between D_1 and D_2.

If, then, the instrument is the interest rate r and its optimal value $\hat{r}$, the two possible values of aggregate demand are given by the points on the horizontal axis corresponding to intersections of the horizontal line at the level of $\hat{r}$ with the curves D_1 and D_2. If, on the other hand, the instrument is the money supply and its optimal value $\widehat{M}$, then the position of the LM curve is so fixed. The two possible values of aggregate demand are, in this case, the horizontal coordinates of the points where this curve intersects D_1 and D_2. Figure 10 clearly shows that, since the $L = \widehat{M}$ curve is upward sloping, the two values obtained in this way will be less far apart from $\hat{y}$ than when the interest rate is fixed at $\hat{r}$. Controlling the money supply is thus preferable to controlling the interest rate.

We understand why such is the case: control of the money supply means that the demand for goods is automatically stimulated in case of an unfavourable shock and tightened in case of a favourable one. This is due to the induced changes in the interest rate because, when the demand for goods is weak, so is the demand for money.

But the conclusions are absolutely the opposite *if random shocks cannot affect demand for goods and services, but rather only to the demand for money*[172]. There is in this case, one single position D for the IS curve and two possible levels L_1 and L_2 for the money demand function, hence two possible LM curves corresponding to the optimal supply $\widehat{M}$ when this is the monetary policy instrument. Controlling the interest rate is thus obviously preferable, since it permits the optimal value $\hat{y}$ to be attained for sure, as Figure 11 shows.

3. It is interesting to make this argument precise by considering the equations supposed to show how the level of aggregate demand is determined. This exercise will suggest, in particular, how a more complete theoretical discussion could be developed.

If we again consider the equations of Chapter 7 and retain only what matters here, linearize them and add random shocks, we can write, for the equilibrium of the goods market:

$$a_y y = A - a_r r + u \tag{320}$$

and for that of the money market:

$$b_y y = M - M_0 - b_r r - v. \tag{321}$$

[172] We know from Walras' law, that the adverb "only" is not really correct, both in this case and in the previous one. We implicitly assume that in both cases there are offsetting shocks on the demand for bonds.

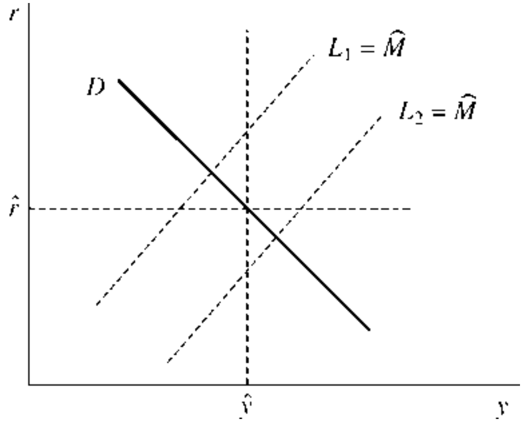

Fig. 11.

In these equations A and M_0 are suitable constants, a_y, a_r, b_y are coefficients whose values were determined in Chapter 7, with:

$$a_y > 0, \qquad a_r > 0, \qquad b_y > 0, \qquad b_r < 0. \tag{322}$$

Finally, u and v are two stochastic terms for which we assume:

$$Eu = Ev = 0, \tag{323}$$

$$Eu^2 = \sigma_u^2, \qquad Ev^2 = \sigma_v^2, \qquad Euv = \sigma_{uv}. \tag{324}$$

A positive value of u corresponds to a higher value than forecast for the demand for goods, a positive value of v to a higher value for the demand for money.

The aim of monetary policy is interpreted as to minimize:

$$V = E(y - \hat{y})^2. \tag{325}$$

Within the present aggregate-demand-analysis framework, this looks like a sensible and simple objective function. If the control concerns the interest rate, we can immediately determine the value of r to be chosen:

$$\hat{r} = \frac{A - a_y \hat{y}}{a_r} \tag{326}$$

hence, the objective function V takes the value:

$$\widehat{V}_r = \sigma_u^2 / a_y^2. \tag{327}$$

If the money supply is controlled, the calculation is a little more involved, but we can make use of the property stipulating that, in this “linear-quadratic” case, the system with no random terms is a certainty equivalent for calculating the optimal decision:

$$\widehat{M} = M_0 + \left[b_y - \frac{b_r a_y}{a_r} \right] \hat{y} + \frac{b_r}{a_r} A. \tag{328}$$

From which we deduce:

$$y - \hat{y} = \frac{-(a_r v + b_r u)}{a_r b_y - b_r a_y} \tag{329}$$

and the corresponding value of the objective function:

$$\widehat{V}_M = (a_r b_y - b_r a_y)^{-2} \left(a_r^2 \sigma_v^2 + b_r^2 \sigma_u^2 - 2 a_r b_r \sigma_{uv} \right). \tag{330}$$

Comparing $\widehat{V}_r$ and $\widehat{V}_M$ we see when it is better to control the interest rate or the money supply. If, for instance, there is no correlation between the two risks u and v (so $\sigma_{uv} = 0$), it is better to control the interest rate precisely when:

$$\frac{\sigma_v^2}{\sigma_u^2} > \frac{b_y^2}{a_y^2} \left[1 - 2 \frac{b_r a_y}{a_r b_y} \right]. \tag{331}$$

The inequality shows that two kinds of characteristics are involved: the randomness of the shocks and the elasticities of the two demand functions. Control of the interest rate appears preferable if shocks on the demand for money are particularly important unless simultaneously the demand for goods is much less sensitive than the demand for money to variations in the interest rate or to variations in output.

4. The main point of the simple model defined by (320) and (321), and of its use here, was to show the importance, for monetary policy, of a proper recognition of the stochastic process ruling the shocks. But once the nature of the approach has been realized, we are inclined to look deeper into the problem that was raised. This was done in several directions, which will be briefly mentioned here.

In the first place, there seems to be no reason why the monetary authority should control either the interest rate or the quantity money. Already in its initial article W. Poole noted that, observing both r and M, the central bank could adjust its interventions in such a way as to aim at a target for a function such as $\varphi = \lambda r + (1 - \lambda)M$, with two positive weights λ and $1 - \lambda$, which would define the policy rule. Instead of being given by (326) or (328) the rule would then be defined by the target $\widehat{\varphi}$ which would minimize the objective (329), i.e. the expected value of the squared deviation from the optimal output. The minimum so reached $\widehat{V}_\varphi$ would then be a function not only of the structural coefficients $(a_y, b_y, \ldots)$ and of the distribution of the shocks $(\sigma_u, \sigma_v, \sigma_{uv})$ but also of the value of the weight λ. A particular choice of this weight would minimize $\widehat{V}_\varphi$ for given values of the other parameters. The resulting optimal "mixed rule" would then improve upon the two rules considered above, which took a single "instrument" into account, either the interest rate or the money supply. Instead of keeping a watch on either the interest rate or the quantity of money, the bank would look at both simultaneously.

In the second place, it was natural to introduce less simple models of the economy and to see how the policy rule depended on the specification. Two extra dimensions immediately came to mind: dynamic models instead of a static one, models that would distinguish between nominal and real variables.

Intertemporal dependence requires simultaneous consideration of successive values for some variables, current output for instance being a function of not just current interest rate, but also of past output and past interest rate. It also requires consideration of serially correlated shocks. Moreover, a closer look at the information available to the monetary authority is required. Indeed, in the heuristic presentation of the problem of this section, we implicitly assumed the monetary authority was instantaneously informed of the values taken by the interest rate r and by the quantity of money M, but not of those taken by output y: if all three variables were simultaneously observed, the problem would disappear because the shocks u and v would be known, through (320) and (321), at the same time as r and M; the problem would no longer be one of choice under uncertainty and the optimal output $\hat{y}$ could be permanently reached[173]. In the spirit of this section we must assume that the output y_t of period t is observed later than r_t and M_t, and we must specify when.

It so appears that a number of dynamic specifications can be envisaged and that the analytical treatment of the problem is likely to be involved. We shall

[173] This reflection leads us to realize that the application made here of the static model is best suited to a world in which the stochastic process of shocks would operate in continuous time.

not survey the literature here[174], but simply report that the specification and discussion provided by P. Champsaur and J. Mélitz[175] lead to maintain the gist of the message issued by Poole's article.

A three equation dynamic model in which both real output and the price level are endogenously determined (together with for instance the quantity of money if the nominal interest rate is taken as exogenous) has been studied by Th. Sargent and N. Wallace[176]. The dynamic structure is quite simple, but the model offers the extra advantage of introducing the hypothesis of rational price expectations in order to take account of the reactions of the private economy to the choice of the policy rule. Again, Champsaur and Mélitz (op. cit.) show that Poole's message essentially stands[177]: the stochastic properties of the shocks matter for the comparison between various monetary policy rules[178].

7.7. A compendium of stability issues

1. Close to the end of the chapter, before wondering about microeconomic foundations, it is appropriate to look with hindsight at the various stability issues raised and discussed at places throughout Chapters 6 and 7 as well as this one. These issues are important in macroeconomics because whatever source of instability may exist causes concern, even perhaps decision to interfere. These issues often differ, and we must be clear about their respective meaning. We deliberately left some of them pending and announced that they would

[174] See in particular J. Kareken, Th. Muench and N. Wallace, Optimal open market strategy: the use of information variables, *American Economic Review* (March 1973); B. Friedman, Targets, instruments and indicators of monetary policy, *Journal of Monetary Economics* (October 1975).

[175] P. Champsaur and J. Mélitz, Une généralisation du choix optimal des instruments de politique monétaire, *Annales de l'INSEE*, No. 46 (1982).

[176] Th. Sargent and N. Wallace, Rational expectations, the optimal monetary instrument and the optimal money supply rule, *Journal of Political Economy* (April 1975).

[177] Sargent and Wallace had pointed to the problem of selecting a particular solution to the system defining rational-expectation dynamic equilibria; existence of the problem had induced them to discard the interest rate from the feasible policy targets. Champsaur and Mélitz exhibit a well defined interest rate policy rule leading to a uniquely-determined stationary solution and fitting well within the approach defined by Poole.

[178] For a more recent list of references to the literature about the same approach, see the beginning of a long chapter which extends this literature in several directions (more policy instruments, more policy rules, interdependent countries, dynamic stochastic simulations): D. Henderson and W. Mc Kibbin, An assessment of some basic monetary-policy regime pairs: analytical and simulation results from simple multiregion macroeconomic models, in: R. Bryant, P. Hooper and C. Mann (eds.), *Evaluating Policy Regimes: New Research in Empirical Macroeconomics* (The Brookings Institution, Washington, 1993).

be taken up again, in particular after explicit introduction of price expectations. Finally we shall not be in a better position in Chapter 9 to survey the range of stability issues; we now reached the proper stage of comprehensiveness; further disaggregations as will be discussed in the next chapter do not much matter and would rather blur the issues.

Our objective cannot be to gather and exhibit firm conclusions about the stability or instability of real economies. Unfortunately such an objective cannot be reached at present. We shall more modestly aim at recalling the important questions earlier raised, at showing their respective domains of relevance and at stating whatever provisional and partial answers we can now give. So, this section may as well be read as suggesting agendas for future research.

To start with we shall isolate four main meaningful questions:

1 – Does inflation tend to accelerate?

2 – Would the economic system left alone lead to a stable equilibrium?

3 – Would the economic system left alone lead to an equilibrium with a satisfactory degree of utilization of resources?

4 – How would increased price and wage flexibility affect the answers to the preceding questions?

Those are meaningful but imprecise questions. Indeed, a good deal of our previous efforts have been to make them precise in particular contexts. We were so led to further distinguish. At this stage three distinctions are worth spelling out:

(i) Is the formation of supplies and demands taken as exogenous? The spontaneous answer is: “Of course, not”. However, when dealing with question 1 above, we sometimes want to isolate the process of price and wage formation and to look at whether it would imply acceleration of inflation independently of its repercussions on supplies and demands, namely under the assumption of unchanging market slacks and tensions.

(ii) When we speak about “the economic system left alone”, do we mean “under fixed values of policy instruments” or “under unchanging policy rules”? In other words, do we take the value of instruments as being endogenous given the policy rules?

(iii) Do we mean the economic system *stricto sensu*, or the whole socio-economic system?

We must finally pay attention to the kinds of justification given to the answers. There is at least the distinction between answers of the type “empirical evidence suggests that ...” and of the type “our theoretical analysis leads us to believe that ...”.

2. Would inflation tend to accelerate under unchanged market slacks and tensions? The simple and now common answer is to say that it will if and only if the unemployment rate falls short of reaching the NAIRU. The answer is not a pure tautology resulting from the definition of the NAIRU, because it rests on the existence of a simple relation between acceleration of inflation and the level of the unemployment rate. The relation may be derived, as we saw, from simple ideas about short-run market adjustments and about adaptive price expectations. The relation also fares fairly well empirically (Section 6.6).

We realized, however, in Section 6.3, that two possibilities existed for the characterization of the price–wage dynamics and that, in order to strictly validate the assumed relation between inflation and unemployment, they required different references. According to the standard characterization, the NAIRU would depend not only on reference values of variables exogenous to the price–wage dynamics but also on a reference value of the rate of inflation; starting from this reference, inflation would then accelerate or decelerate depending on whether the unemployment rate would move to values smaller or larger than the NAIRU. But in the case of complete (long-run) feedback from prices to wages and from wages to prices, the determinant (235) being nil, no such reference rate of inflation would be required for the definition of the NAIRU. Once noted, the difference, which is minor, will not retain our attention any more.

3. In their fight against inflation policy makers often claimed that, if not abated, inflation would spontaneously accelerate. They were then clearly taking a much broader view than that calling for a comparison between the unemployment rate and the NAIRU: reactions from the whole economic, or even socio-economic, system were meant to be involved.

Actually the claims were not precise, for instance they did not answer the crucial question of knowing from which value of the rate of inflation would spontaneous acceleration be triggered. Few, if any, would have gone as far as predicting it to occur from any positive, however small, value of g_p. But it makes a difference if the threshold for appearance of the phenomenon is 5 per cent per annum rather than, say, 15 or perhaps 50 per cent.

Considering how global the assessment is meant to be, we naturally think of looking for empirical evidence. History indeed contains cases of hyperinflation in which acceleration seemed to be overwhelming. But there are also cases in which the rate of inflation remained long contained within a closed interval at a fairly high level. In order to draw lessons from historical evidence about the question at issue, we face the recurrent difficulty of defining the relevant scenarios providing counterfactuals to what was actually observed. Since the statements are meant to bear on "spontaneous" evolutions, these counterfactuals should be characterized by "passive" economic policies; but defining them and assessing what would have happened with them is obviously open to discussion. However, a systematic historical research of that type might be interesting.

More analytical arguments liable to explain the acceleration were given in this chapter. They concerned the changes in institutions and in the formation of expectations that could follow from experience of a sustained inflation.

As was already written in Section 1.9 discussing the costs of inflation, the spread of the practice of indexing constitutes a spontaneous institutional response to sharp, and therefore erratic, inflation. But this spread naturally affects the short-run macroeconomic dynamics. The more widespread will indexation be, the weaker will be the depressive effects of price rises on aggregate demand (see Section 2.4). The larger will then have to be the increase in prices after a stimulating shock (Section 4.1).

We hinted also at various places that experience of high, and therefore erratic, inflation would lead adaptive price expectations to be less regressive than in normal times (in Equation (41) not only would $\hat{g}$ be higher but also would the parameter a be closer to 1). We then saw in Section 4.1 that non-regressive expectations, together with common hypotheses about the formation of aggregate demand, caused instability of the inflationary process.

These two arguments pointing towards increasing instability do not necessarily mean acceleration of inflation. But the conclusion will follow if we moreover assume an asymmetry in the process and argue that the socio-economic system offers more resistance to decreases in the speed of inflation than to increases. This last argument, as well as the preceding ones, would well deserve empirical tests.

Perhaps also the "spontaneous" reactions of authorities to high inflation should be taken into account. On the one hand, since inflation is everywhere viewed as a bad performance of government policy, restrictive demand management might be systematically favoured and succeed in containing inflation. On the other hand, since restrictive policies and their impact on employment are nowhere popular, governments might rather experiment with price and income policies, which bring some positive results in the short run but may soon turn out to worsen the problem for reasons explained in Section 7.3.

4. Would the economic system left alone lead to a stable equilibrium? To pose and study the question is common in any theoretical work based on a dynamic model. Indeed, the model may have a stationary solution or a simple solution implying a regular nice-looking evolution and the theorist wants to know whether such a solution deserves attention, which would hardly be the

case if it was unstable. This is precisely how the stability issue emerged here in the preceding chapter.

Chapter 7 was devoted mainly to the short-term equilibrium, but we also looked a little beyond. Section 5.3 posed the question of knowing whether spontaneous reactions of prices and wages would be stabilizing and we specified for the purpose how prices and wages would react to excess supplies or demands. On this basis a simple dynamic model was set up in Section 5.4 in order to study the interplay of asset accumulation and nominal evolutions in a moving Keynesian equilibrium. We concluded that the solution of the model giving stationary values to real variables was stable, at least if the decrease in prices induced by the state of generalized excess supply was not too fast. We noted, however, that the impact of price changes on price expectations had been neglected and would be discussed later. We similarly studied in Section 5.5 of Chapter 7 two dynamic models of a moving equilibrium with classical unemployment and again found cases of stability.

Such models in which price expectations are exogenous help us to understand part of the phenomena but have a limited domain of relevance, particularly with respect to the process of inflation. In a sense this chapter corrected the deficiency, since its models have endogenous price expectations, but it provides something different from an extension that would add various forms of price expectations to the models of Chapter 7[179]. Indeed this chapter had its own limitations and especially had a different focus. Two of its features matter in this respect.

In the first place, most of our study of dynamic models selected adaptive price expectations. It was partly for analytical convenience, but also because the alternative of rational price expectations, strictly applied, could not easily fit within sufficiently realistic macroeconomic models. The problem was discussed and explained in Section 4.3, in relation with the multiplicity of solutions to equations defining rational expectations, none of these solutions being there persuasive (the multiplicity was already seen in Section 1.5). Perhaps a persuasive way exists, or will be found, for treating stability issues with rational price expectations. But the author of this book only knows at present hybrid attempts at reaching such a aim, hybrid because either the assumed macroeconomic structure is too far from anything realistic for the study of short-run macroeconomic dynamics, or because the formation of expectations does not adhere to the strict logic of rational expectations.

In the second place, when choosing in Section 2.1 a specification for short-run adjustments, we introduced right from the beginning assumptions implying

[179] For what might be such an extension, see the hints given in pages 95 to 101 of E. Malinvaud, *Profitability and Unemployment*, op. cit.

that any evolution in which the rate of inflation g_t would be constant and equal to the value expected for it would be characterized by a zero value of the indicator d_t of demand pressure and by the normal degree $\hat{q}$ of capacity utilization. This restriction on the functions φ and π appears, for instance, in Figure 1 and in the linearization provided by Equations (36) and (37). In models where the labour market is not explicitly represented the hypothesis fits with what is usually believed about the effect of competition on acceleration or deceleration of inflation. But it rules out existence of a moving equilibrium with permanent excess supply and a constant rate of inflation equal to what is expected. The second of the four questions listed at the beginning of this section then is very similar to the third: once the hypothesis is accepted, the same concern about a possible tendency towards full market clearing underlies both questions. For this reason it is more appropriate to deal with the content of this chapter under the heading of the third question to which we now turn.

5. Would the economic system left alone lead to a stable equilibrium with a satisfactory degree of utilization of resources? The question was already approached in Chapter 6, Part 5, about the positive theory of economic growth. The reference was then a balanced neoclassical growth path implying, almost by definition, an efficient use of resources. The question was to know whether such a balanced growth was stable, in other words, whether the path generated, from any initial conditions, by the sequence of subsequent temporary neoclassical equilibria would tend to the balanced growth. We saw that a substantial literature existed about the issue and that it contained many cases of instability, as well as many cases of stability, which were of course fitting better in the simple neoclassical vision. It would serve no purpose to try to summarize here again the results then discussed.

They concerned models of the long-run evolution of a real economy. Loosely speaking, we may consider the treatment of stability issues in this chapter as concerning the opposite extreme and as so providing a complementary image. Simply summarized, they mean that stability cannot be taken for granted. We first easily reached in Section 4.1 the conclusion that, with adaptive expectations and fairly common hypotheses about the formation of aggregate demand, the inflationary process was instable if expectations were not regressive.

This result is closely connected with arguments early raised in discussions about the relevance of Keynesian theory. The question was then to know whether the depressive effect of generalized excess supply on the price level would not spontaneously increase aggregate demand and so lead again to full employment of resources. As we saw at the end of Section 3.3 of Chapter 7, an initial tendency of prices to decrease was said by Keynesians to generate two types of dynamic depressive effects. The first one, related to the worsening of debt-deflation, is ignored in models of this chapter. But the second one lied precisely in the formation of expectations, agents reacting at the first signs of a fall in prices by anticipating a further fall and postponing their purchases. The argument was first formalized in 1975 by J. Tobin along lines similar to those followed in this chapter (his article was quoted at the end of Section 2.1 here).

The result of Section 4.1 can also be interpreted as showing that the dynamic effect will not be strong enough to be destabilizing if, in relation with the force of the induced impact on aggregate demand (here captured through consumption by the size of the parameter a_2 and through investment by the size of a_4), price expectations are sufficiently regressive, agents then keeping in mind that, after the initial decrease in g_t, a return up to the normal rate of inflation $\hat{g}$ would be likely.

The reference in our statement was to the case of passive fiscal and monetary policies, the nominal interest rate in particular remaining fixed. We did not proceed to a complete survey of results concerning other policy rules.

Our discussion in Section 4.1 however suggested that the interest rate could be the instrument of a successful stabilizing policy that government well informed of the dynamic behaviour of the economy could adopt. In Section 4.2 we found that Friedman's monetary policy rule (fixed growth rate of the money supply) would also achieve stabilization, although less quickly and with the occasional occurrence of secondary disturbing cycles.

Such conclusions were not overturned when we discussed empirical work in Section 5.6, a discussion which, however, showed that much was still to be learned for making conclusions more definite.

The same stability issues could have been taken up again after the labour market had been explicitly introduced in the formalization (Part 6). But the focus was then rather placed on the need for econometric investigations if we wanted to reach a precise knowledge about the dynamic behaviour of the economy, with in particular characterization of the NAIRU. The literature certainly contains many relevant contributions which were ignored here. Others will certainly be published in the future.

6. Lack of price and wage flexibility is often nowadays claimed to be responsible for the poor employment performance exhibited by Western European countries in the 1980s and 1990s. The argument is not formalized but rather based on an intuitive confidence in what flexible market economies would spontaneously achieve. Unfortunately intuition of other economists runs counter to such a confidence. The opposition is not new and could be traced back to the eighteenth century.

For what they are worth in this respect, our dynamic models would support the pessimistic view. In particular we may see that all cases of instability found in Section 4.1 for regressive price expectations become more likely when the parameter k_3 increases. In Equation (37) this parameter is precisely the measure of price flexibility: the degree to which inflation reacts to changes in the indicator of demand pressure. Similarly in Section 5.4 of Chapter 7 dealing with the Keynesian depression under exogenous price expectations, the parameter ω characterizes price flexibility and large values of ω lead to instability, which means increasing depression when demand is depressed.

There exists some literature on this flexibility issue. It will not be surveyed here, but it clearly shows that other approaches than the ones followed in this book may also support the pessimistic view. This is for instance the case emerging from a study of De Long and Summers[180] in which wage rates react to rational expectations about output whereas real interest rates react to rational expectations about the price level: increased flexibility in wage rates leads to increased volatility of output. Risk of macroeconomic instability following from price and wage flexibility is also a major theme approached by Hahn and Solow in their recent book[181], particularly in their Chapter 3 and in their conclusion.

The question of knowing whether increased price flexibility would in general be good for the macroeconomy is of course wholly different from that of knowing whether macroeconomic diagnosis cannot detect that, in some countries or regions at some times, real wages were too high because of particular historical developments. Many economists accept the diagnosis that such was the case in Western Europe in the early 1980s and that European unemployment then had "classical" features. Some among these economists think that a quicker correction of real wages would have led to more employment in the following years. It does not follow that the same economists would forever support increased wage flexibility.

[180] J.B. De Long and L.H. Summers, Is increased price flexibility stabilizing?, *American Economic Review* (December 1986).

[181] F. Hahn and R. Solow, *A Critical Essay on Modern Macroeconomic Theory* (MIT Press, Cambridge, MA, 1995).

8 Microeconomic Foundations of Macroeconomic Adjustment Laws

At this point in our book we have to reflect on the importance that adjustment laws assume in the study of the short-term equilibrium or of short-run evolutions. We have to wonder about the foundations of these laws. Let us indeed look back at the development of macroeconomic theory throughout our long eight chapters.

Considering, in Part 3 of Chapter 1, the structure to be given to our models we stressed the concept of equilibrium and the presence of two main types of relationship: the accounting equations, with in particular the consistency requirements of market equilibria, and the behavioural laws, expressing the aggregate effects of decisions taken by individual agents. To obtain behavioural laws the fundamental approach was to first study decisions at the microeconomic level and second derive by aggregation the corresponding macroeconomic relations. At the end of Section 3.6 of that chapter we defined adjustment laws, but they played little role in the subsequent development concerning, first, particular behaviours (Chapters 2 to 4) and, later, economic growth (Chapters 5 and 6).

On the contrary in the preceding chapter and this one we had to carefully specify, early on, the assumed market adjustments. These were discussed in Section 2.1 of Chapter 7, as well as in Sections 2.1 and 6.2 of this chapter. Indeed, the specifications often played a strategic part in defining the properties derived from the models. Whereas we had found appropriate to assume most often that, along economic growth, prices were fully flexible, entailing perfect and continual matches between supplies and corresponding demands, we needed now on the contrary to recognize price rigidities and incomplete market clearing in the short run.

In support of the main specifications of adjustment laws we referred to empirical evidence, based on testimony of market participants or on econometric processing of various data sets. Such evidence will remain the main source of knowledge about such laws in the future. But the fact does not exempt us from looking at explanations derived from a deeper understanding of the phenomena. So doing we may find confirmation of econometric results, or better, we may detect conditions required for their applicability. This will be all the more valuable as we had occasions to conclude from empirical evidence that adjustment laws changed from one period to another, from one country to another. These laws seem to be definitely more dependent on the "institutional context" than are behavioural equations. The study of their microeconomic foundations is then likely to reveal their domains of validity and to point econometric investigations towards significant goals.

The relevance of such a study is well realized. It even happened to be overemphasized when some argued as if price rigidities should be ignored in economic theory as long as they would not be proved to follow from rational microeconomic behaviour. We do not need to share such an extreme and unfounded view in order to recognize the importance of the question already raised at the beginning of Section 3.4 of Chapter 7: what is exactly the rationale of incomplete market clearing and incomplete price flexibility? The time has now come to face the question.

As a research area, the microeconomic foundations in question are challenging, because common sense suggests that they involve many features of our complex world. Modelling each feature in isolation can only explain a part of the phenomenon and will therefore fail to give an accurate account of the facts. Moreover, each one of these features appears as a deviation from the abstract ideal world that the core of economic theory has built in order to be able to deal with various subjects. Introducing the deviation in a model often makes it unwieldy.

Trying here to learn about the outcome of works done in this research area, we claim to be exhaustive neither in our coverage by the selection of models to be discussed, nor in our examination of anyone of those. We shall, however, benefit from what was already done earlier in this book. Indeed, we had earlier opportunities to meet some of those foundations. Briefly recalling how we dealt with them and putting them in a somewhat renewed perspective will suffice.

More precisely the following sections will each be devoted to one of the most important underlying concepts: nature of competition in 8.1, adjustment costs in 8.2, acquisition of information in 8.3, search for opportunities in 8.4, complex exchanges in 8.5, social norms and institutions in 8.6, aggregation in 8.7 and 8.8.

8.1. Facing markets and competition

Figure 1 served in Chapter 7 for the presentation of our hypotheses concerning the short-term equilibrium. It was couched directly in macroeconomic terms with in particular the productive capacity $\bar{y}$. But understanding the behaviour of a representative firm goes a long way in order to explain the graph. So we naturally go back to Chapter 4, and more particularly to the short-term decisions, concerning output and selling prices (see in particular the summary of the two first sections, given at the end of Section 1.2).

The most natural simple reference then is the firm facing a perceived demand curve and acting as a monopoly. Already in Section 1.1 where the simplest assumption about the technology was made, we distinguished two cases as to the effect of an upward (say) shift of the demand function: if the firm's productive capacity was already fully used, just the price of output was raised; if not, both price and quantity increased, but the respective importance of the price increase depended not only on the mark-up (a high mark-up, i.e. a small elasticity of demand, eliciting a small price response) but also on the nature of the shift in demand. Figure 3 indeed opposed two cases, depending on whether the shift meant an increase in outlet (curve C_2) or just an increase in the readiness of buyers to pay a higher price for the same quantity (curve C_1); in the first case the output response was important. Remembering that point P_0 in Figure1 of Chapter 7 was not meant to refer to the situation observed in the preceding period, but rather to the situation expected for the current period, we realize that curve C_1 may not often be relevant for our present analysis of the adjustment process, because it rather pictures what happens in the case of expected inflation.

So, the right part of Figure 1 in Chapter 7 may be said to apply to the case when many firms are bound by their existing capacities. How about the case of tight labour markets, which also corresponds to a high degree of utilization of aggregate productive capacity? Again Section 1.1 of Chapter 4 easily rationalizes this case, if we can accept the hypothesis that the firm freely chooses the wage to be paid to its employees. When labour markets are tight the labour cost function $W(L)$ of the representative firm has a high second derivative: the elasticity of the wage rate with respect to the demand for labour is high. The first of the two Equations (6) of Chapter 4 then shows that the output response to an increase in the demand for the product is then small.

Such monopoly and monopsony behaviours, even extended to long-term behaviours under monopolistic competition, do not fully take account of the actual forms of competition (notice, however, that they have behaviour under perfect competition as a limit when the relevant price and wage elasticities increase indefinitely). The discussion in Section 1.1 of Chapter 4 makes indeed reference to oligopolistic competition and to its implications for the relevance of the concept of kinked demand curves. But obviously the research agenda on this aspect of adjustments to changing circumstances cannot be considered as having been completed.

Notice, however, that we also dealt with wage bargaining in Section 2.2 of Chapter 4. A reason for nominal wage rigidity emerged from the study of the "right-to-manage" model, when we found a case in which the wage rate appearing as the outcome of the bargain depended neither on the labour supply nor on the price at which the firm would sell its output: in this case where some unemployment existed, a rise in the labour supply would lead to an equal rise in unemployment and a rise in the price of output would lead only to increased output and employment. This case was special, not only because the outcome was assumed to be predicted by the generalized Nash solution of the bargaining game, but also because the short-run production function exhibited a constant elasticity of output with respect to the labour input. The case was special but of course revealing, because we chose the assumptions in order to consider a convenient and admissible approximation, not in order to find wage rigidity. Once more, actual forms of competition may lead to less flexible prices than is suggested in textbooks on microeconomic theory.

In Section 8.6 we shall add a few other comments about this delicate choice of a specification for the laws ruling the short-run outcome of wage bargaining.

8.2. Coping with adjustment costs

Changes in prices and in the pace of production entail costs, which may matter even when and where they are fairly small. Their importance actually varies a good deal from one decision to another, from one firm to another, depending on the technologies and the markets involved. Such costs may explain lags in adjustments and rigidity in prices. They indeed received attention in Chapter 4 on two main occasions, for the characterization of respectively the short-term demand for labour and the demand for investment (they were also held as responsible for production smoothing in the theory of inventories). We shall refer here more precisely to the first of the two, but we should not forget that, at the microeconomic level, the very concept of productive capacity would not be interesting if this capacity could be changed instantaneously by any amount without any other cost than the price of the equipment in the case of a required addition and with full recovery of the price in the case of a diminution. Important adjustment costs are so implicitly present in many macroeconomic arguments where productive capacity plays a part.

Section 1.3 of Chapter 4 took the case of the demand for labour as an example in order to look more generally at the effect of adjustment costs. Three dimensions then played significant parts: the specification of the adjustment cost function, the specification of the profit-loss function and the specification of the stochastic process characterizing future uncertainty. Indeed, whether to adjust or not and how much to adjust depend on a calculus comparing the adjustment cost with the discounted expected value of the loss of profits that will result from the fact that no adjustment or just a partial adjustment was made.

Using a number of assumptions the section led to the conclusion that rational behaviour often implied that adjustments to the otherwise optimal decision were partial rather than complete, that high adjustment costs required slow adjustments and that the actual adjustment depended on a longer-term view about uncertainties. Partial adjustments may be responsible for incomplete market clearing, for instance when suppliers on a market choose a price-output combination that deviates from the demand curve, so accepting an otherwise unwanted movement in their inventories.

But the assumptions accepted for the main argument in that section did not fully do justice to the importance of adjustment costs in short-term behaviour, because they contained the convenient but somewhat unrealistic quadratic specification (36) of the adjustment cost function. If there is a positive fixed cost of adjustment, the function is discontinuous at zero. Continuity of the profit-loss function then implies that small changes will never occur. Changes will rather be postponed until there is a sufficiently large discrepancy between the actual value of the relevant variable and those values which would otherwise be optimal at present and for some time in the future (modellization ought then to be somewhat similar to that used in Section 4.2 of Chapter 2 for the analysis of the transaction motive in the demand for money).

Such is precisely the argument that was made by G. Mankiw[182] in order to explain the price rigidities that were observed (those were briefly surveyed here in Section 3.4 of Chapter 7). He started from the idea that in many cases

[182] G. Mankiw, Small menu costs and large business cycles: a macroeconomic model of monopoly, *Quarterly Journal of Economics* (May 1985).

price revisions had a small but positive fixed cost: the "menu cost". Postponing such price revisions was then proved to be often rational.

Whereas the fixed cost of price revisions is usually small, the same cannot be said about the cost of revisions of wage rates applied in a firm, unless nothing special occurred since the preceding revision and implicit or explicit agreement specifies what the revision should be. In many cases the views of employees as to what a fair revision should be seriously differ from the views of management as to what the firm can afford. The revision then implies a negotiation whose course and outcome are uncertain: disputes and strikes may arise, with large disruptions in cash flows. The uncertain and possibly very large fixed costs then explain why a number of revisions are postponed and why a number of others sticks to changes that are easily accepted by "insiders". As a result, actual wage rates exhibit much more rigidity than market rates would.

8.3. Interpreting information

Some rigidity may also be explained by the difficulty that economic agents may experience in interpreting information about changes they observe in their environment. In many cases firms know how the demand for their products changes, but are not able to disentangle what is special to those products from what is induced by a general movement in prices or in real incomes; these firms may not be able either to accurately forecast how the prices of their inputs, hence their costs, will evolve. Under such circumstances their pricing policy is likely to be more cautious than it would be under perfect information.

The idea was explored in a number of articles published in the late 1970s. We may follow here the presentation made by R. Gordon in a relevant survey, where the aggregate effect of such a pricing policy is simply derived[183].

Under certainty and with the approach followed for instance in Sections 1.1 and 1.2 of Chapter 4, we know how price revisions depend on changes in demand and in costs (see, for instance, Equations (6) of that chapter). We just need to transpose the results to uncertainty with a structure of information similar to the one sketched above, and then to look at the aggregate implications.

Let then suppose that there are n firms ($i = 1, 2, \ldots, n$) and that each firm i experiences small proportional changes a_i and c_i in its revenue and cost functions with respect to the preceding period. Under certainty and profit max-

[183] R. Gordon, Output fluctuations and gradual price adjustment, *Journal of Economic Literature* (June 1981).

imization, the relative increase g_i in the price of output of firm i may be approximated by a weighted mean of a_i and c_i:

$$g_i = \psi_i a_i + (1 - \psi_i)c_i, \tag{332}$$

where the value of the weight ψ_i belongs to the interval (0, 1) and depends on the exact specifications of the revenue and cost functions, as well as on the values of the variables in the preceding period. For dealing with the case of uncertainty, we shall simply, like R. Gordon, rely on "certainty equivalence" and say that g_i is equal to the mathematical expectation of the right-hand member of (332).

More precisely R. Gordon starts from the idea that the firm perfectly knows factors directly affecting its operations and its market, but knows only approximately the much more numerous factors affecting its costs. So, in the transposition of (332), ψ_i and a_i may be taken as known with certainty, whereas c_i must be replaced by its expected value. But this should be the mathematical expectation conditional on all the information then available to firm i. Since a_i and c_i are likely to be correlated, we write:

$$g_i = \psi_i a_i + (1 - \psi_i)E(c_i/a_i), \tag{333}$$

where $E(c_i/a_i)$ is, of course, the expectation of the rate of change in costs of firm i, which knows the rate of change in its revenue.

Correlation between changes a_i and c_i reflect both the general movements of the economy and factors acting specifically on firm i

$$a_i = g_v + \alpha_i, \qquad c_i = g_v + \gamma_i, \tag{334}$$

where g_v is here meant to be the rate of change in aggregate income in value term[184].

Whereas the firm observes a_i, it knows only, apart from this, the probability distributions of g_v, α_i and γ_i. If we assume, for example, that these distributions are normal, independent of each other and such that:

$$E(g_v) = \bar{g}_v, \qquad E(\alpha_i) = E(\gamma_i) = 0 \tag{335}$$

we deduce:

$$E(c_i/a_i) = E(g_v/a_i) = (1 - \theta_i)a_i + \theta_i \bar{g}_v, \tag{336}$$

[184] To make the specification even more natural, we may assume that the whole model applies to deviations from medium-run trends.

where:

$$\theta_i = \frac{\sigma_\alpha^2}{\sigma_v^2 + \sigma_\alpha^2}, \tag{337}$$

σ_v^2 and σ_α^2 being the variances of g_v and α_i, as they are perceived by firm i. The result of these assumptions is that:

$$g_i = (1 - \theta_i + \psi_i \theta_i) a_i + (1 - \psi_i) \theta_i \bar{g}_v \tag{338}$$

which may also be written:

$$g_i - a_i = (1 - \psi_i) \theta_i (\bar{g}_v - a_i). \tag{339}$$

Given the simplicity of this linear formula, it is easy to derive its aggregate implications. In particular, if all the firms have the same size, if all the ψ_i and θ_i have the same values ψ and θ, and if the law of large numbers can be applied to the α_i, we directly derive the following result for the mean rate of price increase:

$$g = (1 - \theta + \psi\theta) g_v + (1 - \psi) \theta \bar{g}_v \tag{340}$$

or:

$$g - \bar{g}_v = (1 - \theta + \psi\theta)(g_v - \bar{g}_v). \tag{341}$$

Interpreting this last expression we note that a part of the aggregate underlying unanticipated inflationary or deflationary shock, $g_v - \bar{g}_v$, is realized in the actual rate of inflation, but only a part. The phenomenon follows from the fact that firms, knowing that their clients are prepared to spend more, or less, do not really know how to differentiate between what is specific to their product market, and what concerns the economy as a whole, and will therefore appear in their costs also.

The part which is realized through a price increase is an increasing function of ψ and a decreasing function of θ. We can, in particular, consider two extreme cases. In an economy which has become very sensitive to erratic movements in the inflation rate, θ is small, since agents tend to interpret any increase in the tendency to spend on the part of their clients as an acceleration in the inflation rate (σ_v is thus high). As a result of (341) changes in the growth rate of nominal incomes are thus transmitted almost entirely into changes in the rate of inflation. In the opposite case of an economy where the nominal values usually increase at close to a steady rate, σ_v is small and θ is close to 1; changes in $g_v - \bar{g}_v$ then affect quantities more than prices.

To interpret the model, we should still take into account the lags which are introduced in the formation of expectations and in the transmission of inflationary shocks, at each step in the path of productive processes from the manufacturing of products delivered to final demand, back to raw materials. Nevertheless, this does not fundamentally affect the main result of the model: firms react to changes in their perceived demands according to the way in which they interpret the probable origin of these changes; the more sensitive they are to inflation the more they react by changing prices; but in normal times their price policy is more cautious than it would be under full information.

8.4. Searching before transacting

1. Incomplete information of buyers and sellers also plays in many markets a part which is not recognized in the classical representation where perfectly substitutable items of a given good or supplies of a given service are supposed to be available for anyone ready to pay a given price, and where any new supplier is supposed to be sure to immediately dispose of his or her offer as soon as it is made at the current price. Actually heterogeneity of demands and supplies is often such as to recommend that, before accepting an offer, buyers and sellers proceed on their own to serious investigations about available opportunities. It is the case in particular for real estates or for second hand exchange of durable goods. It is still more crucial on the labour market, particularly so when long attachment to a job is in question.

Searching before transacting is indeed part of normal behaviour. It goes along with frequent lags between the decision to transact and its implementation, often after a period during which the offer has been posted at announced terms, and the terms have been discussed with one or several potential transactors. Search and lags entail costs; they contribute to make transactions more sluggish than is assumed in the usual definition of markets. Commonsense suggests that searching behaviour and the conditions under which it proceeds may have significant consequences of their own, which should not be neglected.

This is why search theory has attracted the attention of economists and now contributes to clarify the microeconomic foundations of macroeconomics. We shall not go deep into it in this book, but only consider some of its applications to the labour market, the domain for which indeed economists are most often referring to it[185].

Analysis of search activities requires consideration of new market concepts, not only the quantities exchanged and the price, but also the still unmet standing supplies or demands, and how their stocks are renewed, as well as the statistical distribution of transaction prices. Some of these concepts were introduced in Chapter 7 when we considered market disequilibria, particularly in Sections 3.1 and 4.2. But we now need a more systematic and closer description of flows of people entering into, or exiting out of, the active part of the labour market, also sometimes called the job market.

To start with, we must point to a possible source of confusion. Up to now we viewed the labour market, the supply of labour and the demand for labour as concerning the whole labour force and the whole flow of its services. But when O. Blanchard and L. Katz[186] write "the labor market has a high level of traffic", they rather refer to what was just called here the active part of the labour market, where transitions take place between being out of the labour force, being unemployed or being employed. Similarly "labour market flows" commonly refer to such transitions, moreover often classified in broad categories concerning the reasons for some of them.

Applying the definitions and conventions selected by labour statisticians, analysts may so consider not only the six main flows (entering into employment from unemployment, or from being out of the labour force, entering into unemployment either from employment or from out of the labour force, leaving the labour force from either employment or unemployment) or aggregates of two of these flows (such as "hirings" or "leaving the labour force"), but also "quits", i.e. voluntary separations of employed from employment, or "firings", i.e. involuntary separations.

Modellization of such flows and econometric analysis of their determinants are taking an increasing room in macroeconomics, in order to either give foundations to familiar tools or assess the likely effects of labour market policies. Fully dealing at this stage with this new macroeconomic approach would be too long. We shall discuss it again briefly in the next chapter (see Section 2.9). We now take a preliminary view of its contribution to our understanding of some microeconomic foundations. In this spirit we shall first consider one of

[185] For an early reference see S. Lippman and J. MacCall, The economics of job search: a survey, *Economic Inquiry* (June and September 1976). See also D. Mortensen, Job search and labor market analysis, in: O. Ashenfelter and R. Layard (eds.), *Handbook of Labor Economics* (North-Holland, Amsterdam, 1986).

[186] What we know and do not know about the natural rate of unemployment, *Journal of Economic Perspectives* (Winter 1997).

the two main theoretical justifications of the Beveridge curve, second briefly examine an early explanation of the Phillips curve.

2. When introducing in Section 3.1 of Chapter 7 the Beveridge curve, in relation with the matching function between labour supply and labour demand, we referred briefly to two considerations: the mismatch between the respective compositions of supply and demand, particularly with respect to locations and skills, and at the end of footnote 26, to search activities "to be examined in Chapter 8". Abstracting from any mismatch we are now going to look, in simple terms, at an equilibrium relation implied by the unavoidable existence of search times before most matches on the labour market[187].

Let us consider the two pools of unemployed people, U in number, and of job vacancies V, and their ratios u and v to the labour force N. These pools are now assumed to constitute two homogeneous sets. For simplicity we argue in continuous time and assume a Poisson process[188] for the occurrence of an unemployed person coming across an acceptable job or leaving the labour force (the two cases feeding the flow of those leaving unemployment). This process is characterized by its probability p: between t and $t + dt$ any unemployed person has a probability pdt to leave unemployment; the probability of being still unemployed after a spell of length l is $\exp\{-pl\}$ and the mean duration of completed spells of unemployment is $1/p$. The probability p, that is the Poisson process, of course depends on the tension between supply and demand on the labour market, a point to be considered in a moment.

From the above assumption follows a simple equilibrium equation involving u, p and the rate of entry e into unemployment per unit of time, the rate being measured as a ratio to employment L. The equation holds along any stationary regime where u, p and e remain constant; the equation is a starting point for various types of development in the modellization of labour market flows. It is

[187] For a short discussion of the interplay between mismatch and search in the generation of the Beveridge curve, see Chapter 5 in E. Malinvaud, *Diagnosing Unemployment* (Cambridge University Press, Cambridge, 1994).

[188] The assumption may be disputed. In contradiction with the basic assumption of a Poisson process, observation has systematically shown that the frequency of exiting from unemployment decreases with the length of the spell of time already spent in unemployment (see, for instance, K. Clark and L. Summers, Labor market dynamics and unemployment: a reconsideration, *Brookings Papers on Economic Activity*, No. 1 (1979)). However, this effect can be simply due to the heterogeneity of the population, an heterogeneity that we deliberately neglect here. The observed phenomenon indeed occurs if the heterogeneous population is made of homogeneous subpopulations, each one characterized by its own Poisson process. Moreover it turns out to be difficult for econometricians to identify the effect of heterogeneity on the aggregate unemployment exit rate as a function of the length of unemployment spells (see J. Heckman, Identifying the hand of the past: distinguishing state dependence from heterogeneity, *American Economic Review* (March 1991)).

obtained from equality between the flow of people leaving unemployment, up as a ratio to the labour force, and the flow of people entering unemployment $e(1-u)$ as a ratio to the same denominator. It may be written in two equivalent forms:

$$e(1-u) = up, \quad u = \frac{e}{e+p}. \tag{342}$$

Implications of this equilibrium equation depend on what is assumed about the determinants of the two rates e and p of entry into, and exit from, unemployment. Analysis of data on labour market flows now provides good information on how the rates vary with various kinds of changes, in particular with changes in labour market tightness[189]. In the literature the standard indicator of this tightness now is the ratio between the vacancy and unemployment rates:

$$\theta = \frac{v}{u}. \tag{343}$$

The entry rate e depends partly on demographic factors and activity behaviour (entry of youth or return of women to the job market). It also depends on institutional changes such as a more common use of temporary contracts which forces some people to search for jobs more frequently. Finally it depends on business conditions and on the corresponding labour market tightness. A downturn in the economy is often characterized by an increase in the number of lay-offs, which is only partially compensated by a decrease in the number of quits. We may therefore write e as a decreasing function of θ and as an increasing function of an exogenous variable z_1 meant to capture the net effect of demographic and institutional factors: $e(\theta, z_1)$.

Observation shows that, on the contrary, the exit rate p is a slowly increasing function of labour market tightness θ. It also depends on other factors captured here by an exogenous variable z_2, so that we can write $p(\theta, z_2)$. These factors are mainly the intensity of search efforts and the eagerness to quickly accept opportunities; they concern employers and unemployed; for the latter they are often represented in the literature by a “reservation wage”, which may react to the level of unemployment benefit. There are also independent institutional changes such as the more frequent use by employers of contracts with limited duration.

[189] O. Blanchard and P. Diamond, The cyclical behavior of gross flows of workers in the United States, *Brookings Papers on Economic Activity*, No. 2 (1990); S. Davis and J. Haltiwanger, Gross job creation and destruction: microeconomic evidence and macroeconomic implications, *NBER Macroeconomic Annual* (1990).

The Beveridge curve was obtained in Section 3.1 of Chapter 7 from the matching function so as to exhibit, between u and v, a relation in which labour market tightness did not appear. Such a relation may be obtained here from elimination of θ between Equation (343) and Equation (342) where e and p are replaced by $e(\theta, z_1)$ and $p(\theta, z_2)$. It is then an equilibrium relation between u and v in stationary regimes, a relation implying that the position of the uv-curve depends on exogenous variables identified by the empirical study of labour market flows.

The analytic form of the relation so resulting from elimination of θ is unwieldy. But the differential of the relation for infinitesimal changes $\mathrm{d}u$, $\mathrm{d}v$, $\mathrm{d}z_1$ and dz_2 is easily found to be:

$$a\,\mathrm{d}u + b\,\mathrm{d}v = (1-u)e'_z\,\mathrm{d}z_1 - up'_z\,\mathrm{d}z_2, \tag{344}$$

where:

$$a = e\big[u + (1-u)\varepsilon_e\big] + pu(1-\varepsilon_p), \tag{345}$$

$$b = up'_\theta - (1-u)e'_\theta \tag{346}$$

with ε_e and ε_p denoting the elasticities, with respect to θ, of e and p. The derivatives p'_θ and e'_θ being respectively positive and negative, the coefficient b in (344) is clearly positive. Observation implies that coefficient a should also be positive, although the signs of the two elasticities point in the reverse direction in (345). Equation (344) shows that the "Beveridge curve" so found shifts to the right when z_1 increases or z_2 decreases, as, for instance, when the participation of women in the labour force increases or when search intensities decrease. Such comparative statics results agree with those announced in Chapter 7.

This definition of the Beveridge curve has, however, the problem of being derived from the analysis of stationary regimes. The speed of convergence towards such a regime is of course rather high, given that the unemployed population gets renewed rather quickly. Nevertheless, we can be almost certain that the curve determined from the comparison of stationary regimes is not exactly appropriate for the transitory phases of the convergence towards such a regime. We can expect that a sudden deterioration in business conditions will rapidly entail a reduction in job vacancies and an increase in entries into unemployment, whereas the effect on the unemployment rate will be more progressive. Therefore, to start with, the move to the right takes place below the Beveridge curve. By the way, this is compatible with observations made on the basis of available statistical data.

3. Let us now consider the simple model proposed in 1970 by D. Mortensen in order to show how the Phillips curve could result from search activities. This is indeed a good opportunity to refer to a collective book which may be considered as a landmark in the development of the concern for microeconomic foundations of macroeconomics[190].

Reduced here to its bare essentials the model is made of three equations: an accounting equation on labour market flows and two behaviour equations respectively concerning the acceptance by workers of the job offers they meet, and the determination by employers of the wages they announce to pay. For the first equation Mortensen identifies two flows: that of transitions from unemployment to employment and the complementary net flow explaining variations in unemployment (entries into unemployment *minus* exits from unemployment to out of the labour force). The ratio e of this second flow, with employment as its denominator, is taken as exogenous and the speed of variation of the unemployment rate is written as:

$$\dot{u} = e(1-u) - \left[f_0 u + f_2(1-u)\right]\varphi. \tag{347}$$

The second term on the right-hand side corresponds to the first flow and deserves attention.

The square bracket represents the number of job offers examined by the unemployed as a ratio to the labour force, whereas φ is the rate of acceptance, i.e. the proportion of these job offers that are accepted. To understand the expression of the square bracket, we just have to know that Mortensen takes as exogenous four ratios, not only e but also f_0, f_1 and f_2. The ratio f_0, defined with unemployment as it denominator, concerns the flow of job offers examined by people who were unemployed when becoming aware of the offers. The ratios f_1 and f_2, defined with employment as denominator, concern job offers examined by the earlier employed, some of whom quit their job for the purpose (those to whom f_2 refers). We note in passing an important difference with the previous model in this section: since e, f_0, f_1 and f_2 are taken as exogenous, the model neglects whichever effects the tightness of the labour market might have on the rate at which workers become aware of job offers that they decide to examine. Most recent works, by Mortensen and others, recognize such effects.

[190] D. Mortensen, A theory of wage and employment dynamics, in: E. Phelps et al. (eds.), *Microeconomic Foundations of Employment and Inflation Theory* (Norton and Co., New York, 1970).

On the other hand, the model takes into account the fact that the acceptance rate φ changes in terms of the gap between the offered wage and the anticipated wage. At the aggregate level this can be written:

$$\varphi = \varphi\left(g_w - g_w^e, \sigma^2\right), \tag{348}$$

where g_w represents the average rate of wage increase and g_w^e the expected value of this rate by workers considering job offers, whereas σ^2 represents the variance of the distribution of announced wages in the different job offers.

To justify (348) Mortensen proceeds with great precision to a microeconomic study. From this he deduces that φ is an increasing and concave function in its first argument. This is because, under the conditions specified, individuals have to make it a rule to accept an offer if the announced wage is higher than a certain level, their "reservation wage", which these searching individuals have determined in advance in terms of the different elements of the model and especially of the frequencies f_0, f_1 and f_2. Now, this level depends directly on what they expect the actual average wage level to be. If they expect too much, that is, if g_w^e is higher than g_w, then the rate of acceptance will be low. We also note that, the slower people are to learn what the exact wage level is, the higher the derivative of φ with respect to $g_w - g_w^e$. A precise study moreover shows that φ has a negative derivative with respect to σ^2: the greater the range of wages proposed, the lower the mean rate of acceptance.

Mortensen also establishes the following equation for average wage offer[191]:

$$g_w = g_p^e + \gamma(w/p, u). \tag{349}$$

To reach this result, he analyses the behaviour of employers. Once they know the attitude of those who contact them for a job, these employers try to get the best of the bargain; this means making a trade-off between different costs and opportunity costs. The higher the real wage, the less interest they have in attracting new workers, which explains why γ has a negative derivative with respect to the first argument. But, Mortensen argues that the higher the rate of unemployment the more this shows that actual employment is low compared with what firms want to have over a longer period. In order to attract workers they would announce higher wages; the derivative of γ with respect to u is so found to be positive. (This comes from the assumption of an overall balance between manpower resources and the medium-run labour needs of firms. This

[191] Remembering the wage adjustment laws studied in Part 6, we note a similarity of the equation with those of these laws which contained an error correction term, more precisely with (232) where adaptive expectations are assumed and into which a term like (238) would be added.

suggests that Mortensen really meant to discuss frictional unemployment in his contribution[192].)

From the exogenous data of the past history of prices and expectations g_p^e and g_w^e the two equations (347) and (349) determine the time path of wages and unemployment. Now, Mortensen shows that these paths tend to go in opposite directions under some commonly occurring circumstances. For instance, if price inflation accelerates, in fact as well as in expectations, Equation (349) shows that the growth rate of wages increases, then according to (348) the acceptance rate φ increases and unemployment decreases: hence, the paths trace a sort of Phillips curve.

It is clear, however, that not all exogenous changes would have this effect within the arguments of the model. For example, if workers' wage expectations were raised, the first effect would be to lower the acceptance rate and increase unemployment. To satisfy their labour needs, firms would have to raise wages. There would be a simultaneous increase in unemployment and in the rate of wage inflation.

This model obviously has other implications, the following, for example: since the reservation wage is a decreasing function of search cost, including the opportunity cost of not taking a job, higher unemployment allowances, which make this opportunity cost lighter, lead to a higher reservation wage, hence to greater unemployment.

The model also attracts our attention to the question of knowing how the expected rate of growth of wages g_w^e, there taken as exogenous, actually reacts. Wage aspirations of people looking for jobs obviously change as a function of the offers made to them, even of those offers they refuse.

We shall not discuss here longer this model and its intellectual underpinnings in 1970. Neither shall we survey the various theoretical and econometric studies that soon followed. We shall take up again quickly the same issue in Section 2.9 of Chapter 9 then considering the labour-market-flow approach at its recent state of development. The main point now was to introduce the reader to an early application of search theory in macroeconomics.

4. The common significant feature of the two examples just discussed is to stimulate a serious study of some microeconomic constraints and behaviours which play an important part in shaping the adjustments as they are observed at the macroeconomic level. Discussions about the exact specifications of functions like $e(\theta, z_1)$, $p(\theta, z_2)$, $\varphi(g_w - g_w^e, \sigma^2)$, $\gamma(w/p, u)$ are obviously relevant

192 The sign so selected for the derivative of γ with respect to u is contrary to what is suggested by our analysis of the wage adjustment law and by econometric fits of this law, such as that given by (274). But a negative derivative would of course make contravariant movements of g_w and u still more likely.

for a deeper understanding of the various adjustment laws specified earlier in this chapter. They draw our attention to the multiplicity of considerations which naturally arise when we so look at foundations. They lead us to realize why macroeconomic adjustment laws may be more fragile and changing than are macroeconomic behavioural laws. So many things may be hidden behind the seemingly innocuous statement saying that adjustment laws depend on the "institutional context"!

It is, however, noteworthy that fundamentally the approach adopted in this section is not different from that used throughout this book. We explore the effects of individual constraints and behaviours, always trying to link the macroeconomic and microeconomic viewpoints. The basic constraints and behaviours may be more interdependent now than they were when we studied, say, the consumption function. But the methodological inspiration is the same.

8.5. Long-term contracts

It is now well recognized that actual price systems are very incomplete with respect to what has to be imagined for a direct applicability of neoclassical general equilibrium theory, as it was developed for fully taking time and uncertainty into account. Instead of exchanges on markets for contingent dated commodities we rather have, on the one hand, exchanges on financial markets concerning a fairly small number of assets giving rights to returns at future dates, depending on the contingencies that will then materialize, on the other hand, long-term contracts agreed between parties engaged in a series of mutual exchanges which they intend to go on for some time into the future. Such long-term contracts are very frequent for exchanges of goods or services, including direct labour services. They explain why prices and wages are less flexible than they would be under the ideal conditions imagined by general equilibrium theory.

Long-term contracts may be fully implicit, as in the "customers markets" already referred to in Section 3.4 of Chapter 7: the seller and its customer are seen as linked by a *bona fide* agreement implying that a "normal price" will be charged. Labour-contracts are explicit, and more or less tightly regulated, but their clauses cannot cover all possible contingencies; although explicit the contracts are incomplete and can be renegotiated at a later date when more will be known. In Chapter 4, Section 2.3, we looked at a simple formalization of such labour contracts subject to renegotiation and saw why they implied less flexible wage rates than those of "outside options" offered by "the market" respectively to the employee and the employer.

We then also examined the microeconomic rational explanations that could be given of the common forms of labour contracts. Then three considerations came into play: the existence of adjustment costs, the nature of the bargaining process, the nature of efficient risk-sharing between the employer and the employee.

Deeper theoretical and factual investigations, into the microeconomic foundations of labour contracts and other long-term contracts for the exchange of goods and services, than those reported in Sections 2.3 of Chapter 4 and 3.4 of Chapter 7 would be welcome. At the present stage we cannot say that long-term contracts have precise implications about the form of adjustment laws, beyond the reassuring confirmation that some degree of price and wage rigidity may indeed by rational. Once more the difficulty comes from the multiplicity of factors that ought to be taken into account and from the diversity of conditions at the microeconomic level.

We may venture to suggest that interesting evidence could come from empirical research projects devoted to a detailed scrutiny of samples of long-term contracts. Special attention ought then to be paid to clauses concerning current adaptations during the implementation of contracts: which revisions in the prices concerned or in the wage rates would be automatic? Which kinds of renegotiations are explicitly allowed? Which escape clauses, if any, are introduced? Answers to such questions would provide useful clues about the domains of validity of such simple adjustment laws as were discussed throughout this chapter.

8.6. Social norms and social institutions

Any casual description of realities about work, employment, wages or labour relations is likely to use concepts developed outside of economics, notably concepts familiar in sociology or sociopsychology. Except for very marginal and occasional comments, such concepts do not appear in this book. Indeed, there is hardly anything to report about their integration into the scientific apparatus of macroeconomics. Could not, however, some of our macroeconomic observations be explained, at least in part, by broader micro-foundations than those usually taught in economics?

The question is intuitively appealing. It motivated reflections and writings by a few economists. It should not be completely neglected in this part of the chapter. But we are not in a position to make a frontal attack to the major research programme that the question suggests. We have to be more modest. In this spirit we cannot do much better than complementing Section 6.1 of Chapter 7, in which we realized that in some respects unemployment can fruitfully be viewed as a social regulator. To begin with, let us appreciate the importance of the challenge as posed by R. Solow[193].

"One important tradition within economics, perhaps the dominant tradition right now, especially in macroeconomics, holds that in nearly all respects the labor market is just like other markets... Common sense, on the other hand, seems to take it for granted that there is something special about labor as a commodity, and therefore about the labor market too" (p. 3)... "Wage rates and jobs are not exactly like other prices and quantities. They are much more deeply involved in the way people see themselves, think about their social status, and evaluate whether they are getting a fair shake out of society... It does not follow... that the ordinary forces of supply and demand are irrelevant to the labor market, or that we can do without the textbook apparatus. It only follows that [this apparatus is] incomplete and need completing" (p. 22).

In his little book Solow selects two main aspects as particularly worthy of consideration: first and foremost, the pervasive part played by equity and fairness in labour relations, second, the delicate dialectics between cooperation and confrontation in negotiations between labour and management. Labour economists, who are used to interchanges with their colleagues labour sociologists, add other dimensions which should not be ignored in a full response to the challenge.

1. An essential point in Solow's book is indeed that we cannot think correctly about the labour market if we do not reflect on how the notion of fairness enters into labour relations. Let us then pause on such concepts as a fair wage paid to an employee, as fair conditions of work, as a fair effort provided by the employee when at work. What do such concept mean? How are they determined in practice? The clue to the answers is the recognition of the precept that you should not take unfair advantage of a favourable position that you happen to be holding.

The notion goes very far in the past. The concept of a just price was much discussed already by scholastic philosophers in the Middle Ages. For St. Thomas Aquinas the just price for whoever was selling, perhaps in a quite thin market considering the conditions of the time, was the normal price holding more generally, something like a broader market price[194]. The main principle behind this identification was that, under normal cost conditions, the seller should not charge an exceptionally high price when he or she could do so, because the buyer would have a particularly urgent need of the good to be sold,

[193] R. Solow, *The Labor Market as a Social Institution* (Basil Blackwell, Oxford, 1990).

[194] See D. Friedman, Just price, entry in: J. Eatwell et al. (eds.), *The New Palgrave Dictionary of Economics* (1987), op. cit.

or even simply would particularly benefit from holding the good. Similarly Solow quotes A. Marshall writing: an employer "acts unfairly if he endeavours to make his profits not so much by able and energetic management of his business as by paying his labour at a lower rate than his competitors; if he takes advantage of the necessities of individual workers, and perhaps of their ignorance of what is going on elsewhere".

Solow adds "Marshall is not claiming to describe the way wages are actually determined in England in the 1880s... [but] he thinks that the tendency of men to want and employers to offer fair treatment plays a significant part in the operation of the labor market then and there. He also approves of it". That all this conforms with popular conceptions of fair wages and prices was demonstrated by D. Kahneman, J. Knetsch and R. Thaler[195]. That the *social norms* following from these conceptions play a significant part in the actual determination of wages and prices is also an important message of the literature we are now considering.

The normal wage that it is fair to offer assumes the existence of a reference. The three authors just mentioned, as well as G. Akerlof and J. Yellen[196], persuasively argue that this is not quite a market-clearing wage rate. The positive correlation often found between profit and wages, for instance, across industries, shows that employees are in a sense also shareholders: their wage rate reflects the good or bad fortunes of the industry in which they are employed, even those of their firm. Still more revealing is the fact that all workers in better-paying industries tend to receive positive wage premia: the wages of secretaries and engineers are highly correlated across industries.

The same concept of fairness implies that in weighing the intensity of their effort employees should not take advantage of the fact that effort happens not to be observable or observed. Again the expected normal level of effort should be provided, at least to the extent that a fair wage is being received. This precept also is very much rooted in common psychology. Solow points to the fact that, in cases when welfare recipients were asked to provide some work in exchange, a large majority of them reported that they were satisfied to receive their benefits tied to a job, as compared with just receiving the benefits.

We are so led to associate reciprocity with the notion of fairness, also called equity. Indeed, the importance of reciprocity in many human relations was stressed by sociologists when they spoke of "the exchange of gifts". It is also taken for granted by the teaching in business management, according to which

[195] D. Kahneman, J. Knetsch and R. Thaler, Fairness as a constraint on profit seeking: entitlements in the market, *American Economic Review* (September 1986).

[196] G. Akerlof and J. Yellen, Fairness and unemployment, *American Economic Review* (May 1988).

the most important aspect in the framing of a compensation system is its accordance with workers' conceptions of equity. This widespread evidence should be a warning to economists when they draw extreme consequences of self-interest or when they give a too narrow interpretation of what self-interest means. Such was the point made by G. Akerlof when he argued that paying "efficiency wages" was in the interest of employers even independently of any fear that some employees could quit and of any punitive treatment to those employees who would be shirking[197]. Such is also the main outcome of interviews of business managers. To the references and conclusions reported at the end of Section 2.4 of Chapter 4 we may now add the full presentation of the extensive set of interviews led by T. Bewley[198]: of all rationalizations invoked by economists in order to explain wage rigidity, the ones referring to equity and workers' morale were by far the most frequently accepted by managers. The importance of reciprocity is linked with the role of mutual *trust*, which is commonly recognized as favourable to a smooth operation of our economies.

Reciprocity was also perceived by A. Marshall, as quoted by Solow again for the case in which product demand increases: "As a general rule employers will be bound in fairness to yield at once ... a considerable part of their new profits in higher wages", without waiting to be compelled to do so by industrial action of any kind. In the other direction: "Fairness requires a similar moderation on the part of the employed". Employers are not supposed either to take advantage of increases in their own bargaining power that come about because employment is irregular. Solow adds bits and pieces of modern evidence showing that the moderation praised by Marshall is still at work.

Overall, the importance of equity and trust, and of social norms about them, leads us to realize the existence of psychological reasons for a good deal of what we observe on labour markets. These reasons must be recognized in what we call the micro-foundations of macroeconomics. They reinforce in particular the validity of the efficiency wage hypothesis discussed in Section 6.1 of Chapter 7.

Undoubtedly we could find similar, although less pervasive, considerations to present about other markets. Even for labour markets the part played by sociopsychology is not limited to the formation of wages. For instance, the persistence of unemployment, and its polarization in some regions, even in some families, seems to have something to do with the weakening of the norm of employment (i.e. of the precept "you should be employed") whenever and wherever unemployment becomes a more familiar experience. Indeed, in his

[197] G. Akerlof, Labor contracts as partial gift exchange, *Quarterly Journal of Economics* (November 1982).

[198] T. Bewley, Why not cut pay?, *European Economic Review* (May 1998).

study of the British Household Panel Survey, A. Clark[199] finds that unemployment of "relevant others" makes the unemployed less unhappy, less likely to search for a new job and less likely to be reemployed at a later date, other things being equal.

2. Other micro-foundations may be inspired less by psychology and more by the sociology of labour relations. First of all, other social norms than those derived from the sense of equity may come into play, such as: "you should not undercut wages", i.e. steal the jobs from an employed worker by accepting to work in his stead at a lower wage, or "you should not break a strike", i.e. accept to work when your fellow workers are on strike. The sense of class solidarity behind these precepts was instilled by a long history of industrial relations.

In their presentation of the insider–outsider theory A. Lindbeck and D. Snower[200] convincingly argue that it is in the interest of both the insiders and their employer to abide by the precept of no-wage-undercutting. "Insiders may be more cooperative and friendly to one another than to underbidding entrants in order to prevent [their] wage to be eroded... Firms may be unwilling to hire outsiders at less than the prevailing wage, because, given the insiders' unwillingness to cooperate with underbidders, it is not profitable for the firms to accept underbidding" (overview of their Chapter 5). Employers know the importance of cooperation or harassment: by cooperation activities workers support one another in the process of production and thereby raise each other's productivities; in contrast, harassment activities are designed to make each other's job more unpleasant and thereby raise each other's disutility of work and damage overall productivity; outsiders then become victims of discrimination. The two authors present their economic analysis of the situation so described and show that their results have roughly the same macroeconomic implications as those following from the existence of hiring and firing costs.

G. Akerlof[201] similarly argued that other features of labour relations had macroeconomic implications akin to those announced by efficiency-wage models. He pointed in particular to the sociological concept of segmented labour markets and to its economic analysis by P. Doeringer and M. Piore[202]. These authors simplify reality by saying that there are two types of jobs, re-

[199] A. Clark, Unemployment as a social norm: psychological evidence from panel data, LEO-CRESEP, Université d'Orléans (December 1998).

[200] A. Lindbeck and D. Snower, *The Insider–Outsider Theory of Employment and Unemployment* (MIT Press, 1988).

[201] G. Akerlof, Gift exchange and efficiency-wage theory, *American Economic Review* (May 1984).

[202] P. Doeringer and M. Piore, *Internal Labor Markets and Manpower Analysis* (Lexington Books, Heath, 1971).

spectively in two segments or "sectors" of the labour market. Primary sector jobs have stability, low quit rates, good working conditions, acquisition of skills, good pay and career prospects. In contrast, secondary sector jobs are often temporary or have high quit rates, harsh discipline, little chance of promotion, low acquisition of skills and poor pay. In order to describe the regulation of promotions and pays within the primary sector, Doeringer and Piore introduce the concept of an internal labour market (internal to the firm). Akerlof argues that the theory then given follows closely the sociological theory of organizations and bureaucracies as developed since Weber. Be that as it may, the difference in pay between primary and secondary sector jobs can be seen as the difference between wages in excess of market clearing and wages at market clearing. Provided firms set the primary sector wages as they prefer, the dual labour-market hypothesis leads to an efficiency-wage theory of the operation of the primary sector where access to jobs is rationed, many workers employed in the secondary sector being of course candidates to primary sector jobs.

Akerlof also points to studies by empirical sociologists observing work groups. The studies show considerable discrepancy between the formal and the actual authority in many different work situations, as well as between the official work rules and the customs applying in practice. Within this framework where adherence to authority is less than complete, the loyalty of employees may much contribute to high productivity. This loyalty is exchanged for high wages; it can then be translated via effective management into high productivity. Such empirical observations so complement the idea that "exchanges of gifts" have a role to play in efficiency-wage theories. This role is clearly more complex than what the notion of reciprocity between fair wage and fair effort would alone suggest.

3. Let us now pause and take stock of the macroeconomic implications of such psychological and sociological considerations as were just reported. These considerations have reinforced the explanations given in Section 6.1 of Chapter 7 addressed to the existence of involuntary unemployment. They did it indirectly when they brought new arguments in favour of the efficiency-wage model or the insider–outsider model. They did it directly also when they bore on the segmentation of the labour market, because this market could not be in equilibrium if the search for primary sector jobs would quickly succeed, thus leaving too many secondary sector jobs vacant.

In this chapter, where we are more interested in macroeconomic dynamics, we naturally wonder how social norms are likely to change in interaction with other changes. Indeed existence of such a dynamic interaction follows from the fact that norms are rooted in the experience of people: when experience changes norms should adapt and also change. This concerns in particular wage rates since we argued that fair wages were defined by reference to normal wages. We then expect that, at the macroeconomic level, the wage norm moves in the direction of the net excess demand for labour, which creates a tension on actual wages. But the flexibility is definitely lower than it would be with market-clearing wages: actual wages lag behind market-clearing wages, and just wages lag still more.

We note in passing that the social norms in question must concern the purchasing power of pay, hence real wages rather than nominal wages. This means that, as soon as people are conscious of living in inflation, norm-induced wage rigidity also concerns real wages. At any horizon, except those too short to be really interesting in macroeconomics, the relevant wage rigidity so bears on real wages.

Wage rigidity is of course an essential notion in the specifications discussed in this chapter, specifications which precisely aimed at measuring speeds of adaptations, hence degrees of flexibility. It is in this respect that social norms and institutions are providing particularly valuable micro-foundations. The point is so obvious that it needs no elaboration. But we may note here that wage rigidity has often been seen as responsible for the persistence of unemployment and that, bypassing the chain of arguments and empirical proofs establishing this responsibility, a number of economists jump to the conclusion that the micro-foundations of unemployment persistence lie in social norms and institutions. As we shall see in Section 2.9 of Chapter 9, the conclusion might well be a bit premature, because our understanding of the persistence in question is still very fuzzy. This is unfortunate because the point as some importance when attention moves from the subject of this chapter, short-run dynamics, to the theory of growth: should we say that the micro-foundations discussed above play little role in the long run in comparison with the determinants commonly identified in growth theory? Probably not if they cause unemployment persistence.

4. Studying the social institutions of the labour market, labour economists often focus on negotiations between workers and managers, with the mix of co-operation and confrontation characterizing them. Macroeconomists certainly have something to learn from this literature. But what exactly?

The answer given in this book is cautious. Roughly speaking we may say that the book presents an analytical and modular methodology, according to which labour market institutions explain some detailed features of adjustment laws but not the overall architecture of the macroeconomic system. We discussed the implications of wage bargaining in Section 2.2 of Chapter 4, the main outcomes then being meant to concern the labour demand function of firms and the wage adjustment law. In Section 5.6 of Chapter 7 we used the concept of “bargaining equilibrium” in order to easily capture the idea that, given labour market institutions and the behaviour of participants in wage negotiations, a higher employment would lead to a higher real wage, and in order to recognize the possibility of shocks originating in labour relations. But the main part of the model, the “economic equilibrium”, was kept immune from any other interference. Similarly in this chapter the variable z_{wt} of Equation (232) was meant to capture the impact of exogenous factors influencing wage negotiations.

In the second chapter of his book Solow gives a much more crucial role to cooperation or conflict in wage negotiations, which are close to becoming the main single determinant of the macroeconomic equilibrium[203]. Let us briefly look at this part of his analysis.

In order to give a flavour of the logic he is going to apply, Solow first refers to “the prisoner’s dilemma”, which is familiar in the theory of games. There are two players, A and B; each has to choose between two moves, either abide by a cooperation strategy or enter into conflict; let us call then C (for cooperation) and D (for defection); there is no possibility of either preliminary communication or credible binding commitment by a player about which move he will choose. Each player knows that cooperation (C, C) is better for both than conflict (D, D), and each player also knows that if he chooses C and his opponent chooses D, he will be worse off than with (C, C) and his opponent will be better off, and vice versa.

Now, the interesting framework for Solow is one of repeated games, in which moves are chosen by the same players period after period. Each player in the prisoner’s dilemma game then chooses his move in period t as a function of all previous moves by both: specifying the function is so defining a strategy. The interesting concept in the theory of such repeated games with two players is that of equilibrium couple of strategies (or equilibrium for short) in which the strategy of a player is rational for him given the strategy of his opponent. Of course for a rigorous theory the repeated game and the equilibrium concept must be more precisely defined. But two types of result frequently appear in this theory and are interesting with respect to the argument exposed by Solow: permanent cooperation turns out to be an equilibrium; many other equilibria also exist.

Solow considers a model of the labour market game in which cooperation means that the wage rate is at the level of a social norm, but at this level employment, decided by employers, is too low: permanent cooperation means

[203] In retrospect we may wonder whether this is not a message that Solow has long aimed at sending to macroeconomists. The message can be read in some of his earlier works (see in particular his two articles with I. Mc Donald, quoted here in Chapter 4).

permanent unemployment. It is cooperation on the part of those workers who happen to be unemployed, because each one of them would have the possibility of undercutting the wage rate and to become employed. But undercutting would trigger competition among workers, so that the wage rate would fall up to the level of the reservation wage, where all workers would be indifferent between being employed or not. For some values of the parameters of the model, even the workers who undercut would be worse off in comparison with what would have happened to them waiting until the time of their chance to be employed. In other words, a range of values of the wage rate would exist for which no-undercutting would be rational as long as nobody would undercut. Within this range the higher the wage rate, the more unemployment there would be.

Solow makes it explicit that his labour market game "is intended as an extended metaphor, not a literal description". We should rather understand that experience and reason of workers and employers lead to the emergence of social norms: for the fair wage, for the no-undercutting behaviour and for the acceptance of unemployment, so long as there is not too much of it.

If their Chapter 5, F. Hahn and R. Solow[204] come back on this type of argument with other models of the labour market. They claim that their models, metaphorical as they may be, contribute to "the realistic approach to the labour market", in contrast with the scholastic market-clearing approach. The author of this book has much sympathy for this attitude with respect to the proper approach to the labour market. On the other hand, he is vary of the present fashion to stop at metaphors in theoretical economic research, even when it is claiming to provide foundations. He would like to know better, for each metaphor, the contour of the domain beyond which transposition to the real world is more misleading than instructive. He usually finds the authors of metaphors, and there are many these days, too silent about his queries.

8.7. Aggregation of adjustment laws

We cannot leave the discussion of microeconomic foundations of adjustment laws without wondering how adaptations decided by individual agents aggregate. The question was casually approached at places in the preceding sections. In each case it could be better studied. Would, however, a more transversal outlook be liable to be informative? This short section is meant to hardly more than raising the issue. It will contain first two comments on the consequences following from aggregation of behaviour, second a brief discussion of aggregation of Phillips curves over micro-markets.

[204] F. Hahn and R. Solow, *A Critical Essay on Modern Macroeconomic Theory* (MIT Press, 1995).

We might first perceive an antinomy between the apparent macroeconomic reactivity of price or wage inflation to market tensions and the often alleged lack of freedom of individual agents in their price or wage policy. Indeed, in most cases firms and trade unions pretend to have little room for manoeuvre. Prices in particular would evolve in proportion to costs pretty much in agreement with the simple Leontief model.

Actually, if we stick to prices only, we have to look at the whole chains of transformation from primary factors to goods and services entering consumption or output. Even if markups would be fixed at most stages, market tensions would still be felt at some of them, particularly for the prices of raw materials. Also small changes in markups tend to have cumulative effects along the chains if market tensions are widespread. Finally such effects need not be large in order to agree with what is observed at the aggregate level. Indeed, market tensions do not play a large part in the econometric fits of the price equation, as we saw in Section 6.4.

But, and this is our second comment, aggregation of behaviours with respect to adjustments is likely to lead to longer lags than those suggested by the study of representative agents. Whether due to fixed adjustment costs, to search or to rigidities introduced by long-term contracts, microeconomic lags also are cumulative along the production process from primary factors to finished products. But they are more or less cumulative, depending on how often firms change the price of their output in anticipation of changes that did not yet materialize in the prices of their inputs. Acting in anticipation may appear more likely in case of rapid inflation because the stakes are more important; so there seems to be good microeconomic reasons for the observation referred to at the end of Section 1.4, according to which lags in adaptive expectations and adaptive behaviours tends to decrease when the rate of inflation increases.

Let us go on now to a brief examination of the aggregation of "Phillips curves". The question, which is obviously representative of a whole category of similar questions, has played, right from the beginning, a role in the discussions which followed the article by A. Phillips[205].

So, consider n labour markets ($i = 1, 2, \ldots, n$). On each market there is a supply N_i of labour and a demand D_i for labour. Employment L_i is equal to the minimum of the two because the labour market is assumed perfectly homogeneous and perfectly efficient in matching demand and supply at a given wage rate. This matching gives the unemployment rate u_i or the vacancy rate

[205] See in particular R. Lipsey, The relation between unemployment and the rate of change of money wage rates in the United Kingdom 1862–1957: a further analysis, *Economica* (February 1960). This article, which considered various aspects of Phillips' equation, was later taken as the starting point for a number of contributions.

v_i depending on the case[206]:

$$L_i = \mathrm{Min}(N_i, D_i), \tag{350}$$

$$u_i = 1 - \frac{D_i}{N_1} \quad \text{and} \quad v_i = 0 \quad \text{if } D_i \leqslant N_i, \tag{351}$$

$$v_i = \frac{D_i}{N_i} - 1 \quad \text{and} \quad u_i = 0 \quad \text{if } D_i > N_i. \tag{352}$$

On each market i a simple sort of Phillips law operates: although the wage rate is not perfectly flexible, its growth rate g_i increases as a function of the ratio of the demand for, over the supply of, labour. We may write:

$$\begin{aligned} g_i &= \psi_i(u_i) && \text{if } u_i > 0, \\ g_i &= \psi_i(-v_i) && \text{if } v_i > 0, \\ g_i &= \psi_i(0) && \text{if } u_i = v_i = 0, \end{aligned} \tag{353}$$

where ψ_i is a decreasing function. This Phillips law can still be considered as expressing the form taken by the "law of supply and demand". Such a function ψ_i is graphed in Figure 12.

In order to reflect on the implications of aggregation over micro-markets, we may focus attention on the case when the same function ψ_i applies to all markets and when these have equal sizes, so that the aggregate rate of wage increase g_w is approximately equal to:

$$g_w = \frac{1}{n} \sum_{i=1}^{n} g_i. \tag{354}$$

Similarly, the aggregate unemployment rate is approximately equal to

$$u = \frac{1}{n} \sum_{i=1}^{n} u_i. \tag{355}$$

For given values of the rates $u_1, u_2, \ldots, u_n$ and $v_1, v_2, \ldots, v_n$ such that, by hypothesis, $u_i v_i = 0$ for all i, Equations (353), (354) and (355) define the coordinates of a point (u, g_w) in a graph such as that of Figure 12. A qualitative discussion on how this point may evolve will suffice for our purpose.

206 For simplicity we no longer distinguish, as we did in Section 3.1 of Chapter 7, between the labour force and the supply of labour, nor between the number of jobs and the demand for labour.

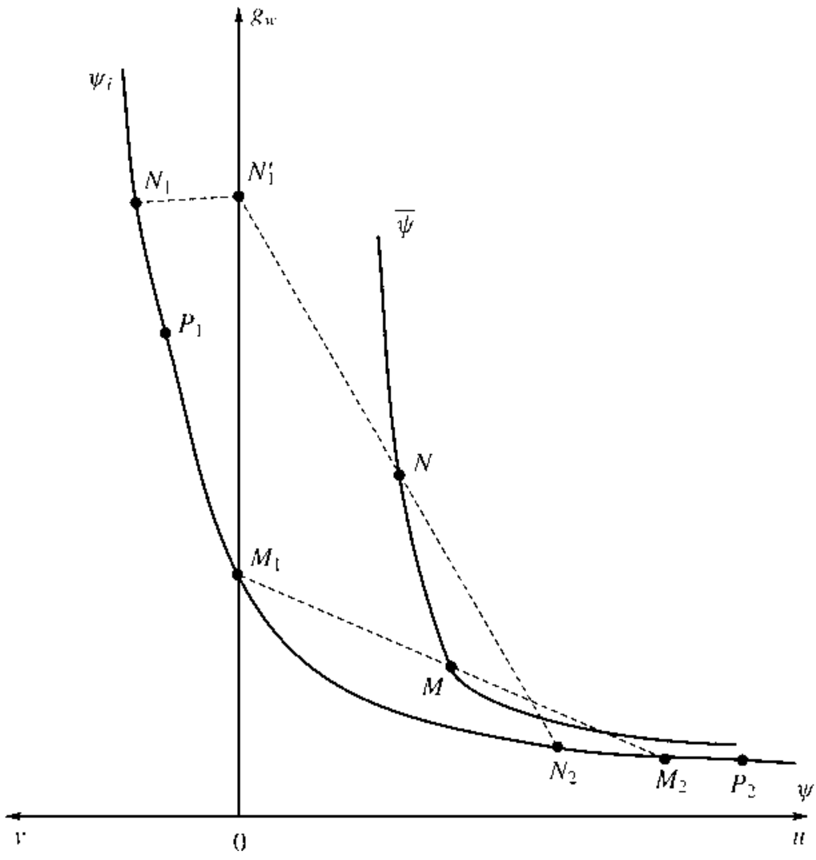

Fig. 12.

Consider the case in which there would be just two micro-markets. We may use the graph in order to find the exact location of (u, g_w). For instance, if the vacancy rate is nil in both micro-markets (u, g_w), is the middle point of the segment defined by the two (u_i, g_i): in the graph the aggregate is given by point M when the situations in the two markets are given by points M_1 and M_2, which belong to curve ψ_i. The construction is hardly less easy when the vacancy rate is positive in one of the markets: given the two positions N_1 and N_2, the aggregate is found as the middle point N of the segment defined by $N_1' N_2$. We may even trace the movement of the aggregate point (u, g_w)

when the points representing situations in the two micro-markets move, the difference between D_1/N_1 and D_2/N_2 remaining however unchanged; we thus obtain a kind of aggregate Phillips curve $\overline{\psi}$.

Two conclusions follow. If ψ_i is upward-concave like in Figure 12, $\overline{\psi}$ also is. Concavity even appears more marked on $\overline{\psi}$ than on ψ_i because of the discontinuity in the slope at point M. If we start from two points, P_1 and P_2 say, such that the difference between D_1/N_1 and D_2/N_2 is larger than with M_1 and M_2, the aggregate point is located above $\overline{\psi}$. In other words, no exact aggregate Phillips curve could be defined independently of the dispersion of situations between micro-markets: for a given unemployment rate u, wage inflation g_w is all the higher as the dispersion is higher[207]. Clearly, these two qualitative conclusions stand as well if we consider more than two micro-markets.

Dependence of wage inflation on the dispersion of the state of market tensions between micro-labour-markets has often been claimed to matter, notably when and where shortage of skilled labour was taken as responsible for an overall acceleration of inflation.

8.8. Aggregation of linear dynamic equations

All through the chapter we used specifications which were explicitly dynamic. We paid particular attention to the structure presented in Section 6.2 with Equations (229), (231) and (232). These dynamic adjustment equations were directly specified at the aggregate level and meant to be, first, tested, second, estimated by econometric fits on macroeconomic data. However, microeconomic ideas were not absent from the considerations leading to the choice of a particular form for a particular equation. We may say for instance that behind Equation (229) stood the notion of a short-run labour-demand function reacting to changes in output demand and exposed to adjustment costs (as explained in Section 1.3 of Chapter 4). Or behind Equation (231) was the idea of price-making firms exposed to monopolist competition and to changes in costs as well as in demand. Then the question naturally comes to mind: should the macroequations be assumed to have the same dynamic properties as directly come out from microeconomic analysis?

The question does not concern only adjustment laws. It applies, with even increased force, to the dynamic behaviour equations more and more often found in the macroeconomics of employment, inflation and business fluctuations. We

[207] Such a conclusion is not surprising: as we saw when considering in general the statistical approach to aggregation (Section 2.3 of Chapter 2), relations between aggregates most often depend on the statistical distribution of characteristics of micro-units.

shall see in the next chapter, specially in our discussion of the models promoted by the real-business-cycle movement, that precise dynamic behaviour equations are directly borrowed from analysis of the behaviour of a representative agent and then play an important part in predicting the dynamics of the economy. This is why we may welcome the research of M. Forni and M. Lippi into the aggregation of linear dynamic relations[208]. This is now the best place for locating it in our book where it should not be overlooked: indeed, it displays an approach and a sample of results which are too relevant to be forgotten.

In the introduction to their book, Forni and Lippi write: "Do the features of the micromodel, as derived from micro theory, survive aggregation? ... If we rule out *ad hoc* assumptions on the distribution of the microparameters over the population, then aggregation spoils any pleasant property of micromodels ... On the other hand, the difference between micro- and macromodels, which arises from aggregation, can be the source of reconciliation between theory and empirical data ... The message of this book is that macroeconomic modelling and testing would receive a new impetus if a better balance were reached between micro theory, aggregation theory, and empirical research on the distribution of the microparameters over the population".

The approach potentially embodies three features which ought to be recognized in modern macroeconomics: heterogeneity of agents, heterogeneity of shocks, specification of what information is available about shocks. In this short presentation we shall give just a few hints on various parts of the analysis.

1. The fact that simple dynamic properties do not survive aggregation over heterogeneous agents is not surprising, because it is quite congruent with what has long been known about the effects of aggregation on linear relations. Here, it will suffice to repeat an example given by Forni and Lippi in their 1999 article. Assume that the decision of agent i in period t makes a variable y_{it} proportional to an independent variable x_{it}, which is a two-period moving average of a white noise u_{it}:

$$y_{it} = a_i x_{it}, \qquad x_{it} = u_{it} + \alpha_i u_{i,t-1} \tag{356}$$

with a positive α_i. Then, except if all a_i are equal, the simple aggregates Y_t and X_t over the entire population will not be proportional. Any model of the form

$$Y_t = AX_t + \Omega_t \tag{357}$$

[208] M. Forni and M. Lippi, *Aggregation and the Microfoundations of Dynamic Macroeconomics* (Oxford University Press, 1997); M. Forni and M. Lippi, Aggregation of linear dynamic microeconomic models, *Journal of Mathematical Economics* (February 1999).

will have to allow for existence of a non-zero additive term, which may be called an "aggregation error" by us and may be taken as a random term by the macroeconomist. For instance, if there are only two micro-units:

$$\Omega_t = (a_1 - A)(u_{1t} + \alpha_1 u_{1,t-1}) + (a_2 - A)(u_{2t} + \alpha_2 u_{2,t-1}). \tag{358}$$

If $a_1 \neq a_2$, Ω_t will have a positive variance and will be positively correlated with Ω_{t-1} and Ω_{t+1}, whatever the choice of A (it is natural to choose A as minimizing the variance of Ω_t). So the exact static relation between x_{it} and y_{it} aggregates into a stochastic relation with a serially correlated disturbance.

The change in the dynamic properties from the micro- to the macro-model is actually more pervasive than this simple example may suggest. Not only may static microequations be transformed by aggregation into dynamic macroequations, but also micro-cointegration does not imply macro-cointegration as we shall see, nor does lack of Granger causality in the micro-model imply the same property at the macro-level. Overidentifying restrictions imposed by the underlying micro-theory cannot be tested directly using aggregate data. Actually, aggregate implications of micro-theory can only be found by explicit modelling of heterogeneity, which is likely to require a lot of additional information with respect to the direct transposition which is common at present in macroeconomics.

2. We realized in this chapter, particularly in Sections 7.4 to 7.6, the importance of a correct characterization of shocks to which the macroeconomy was exposed. As soon as we look for micro-foundations in this respect, we have to recognize that many shocks at the level of micro-units are more important for those units than what comes from macro-shocks. So, to the distinctions between permanent or transitory shocks, and between supply, demand or price shocks, must be added a distinction between those shocks that affect only few individual agents and those affecting many simultaneously. For our reflection the two extremes will suffice: we shall speak of an idiosyncratic shock η_{it} concerning agent i in period t if the shock leaves all other agents unaffected; at the other extreme a "common shock" u_{kt} will concern all agents. The distinction is reminiscent of the one between the two kinds of exogenous variables, which we found to be required right from the beginning of our study of aggregation of economic relations in Section 2.1 of Chapter 2: the components of vector x_i were idiosyncratic, i.e. specific to agent i, those of vector z designated common exogenous variables. The two kinds of variables indeed behaved differently under aggregation.

The distinction is particularly sharp for linear models. As a background for their study of aggregation in dynamic linear models, Forni and Lippi discuss a specification in which the shock ξ_{it} hitting a microeconomic variable x_{it} is made of $h+1$ components:

$$\xi_{it} = a_{i1}(L)u_{1t} + \cdots + a_{ih}(L)u_{ht} + \eta_{it}. \tag{359}$$

In this equation the last term η_{it} is meant to be idiosyncratic, whereas the h random variables u_{kt} are common to all units (with $k = 1, 2, \ldots, h$). The multiplier $a_{ik}(L)$ is a polynomial, or at least a rational function, of the lag operator L. Heterogeneity of agents is reflected not only by the presence of the idiosyncratic component, but also by the fact that the multipliers depend on i, in the same way as does the coefficient a_i of the simple static model (356), in which it was however multiplying the idiosyncratic exogenous variable. Heterogeneity of shocks appears not only with the distinction between the idiosyncratic shock and the common shocks, but also with the multiplicity of the latter.

All components must appear in the analysis of decisions taken by agent i. For instance, if the precautionary motive matters and if the variance of η_{it} dominates in the variance of ξ_{it}, idiosyncratic shocks will play an important part in explaining individual behaviour. With large populations we may expect the idiosyncratic components to cancel out in their effect on the exogenous variable X_t obtained by aggregation of the x_{it}: this is a consequence of the law of large numbers, which applies if the individual η_{it} are mutually independent but applies also under much weaker assumptions. This cancellation is an aggregation effect which simplifies aggregate analysis. However, as we just argued, it does not make idiosyncratic shocks absolutely negligible in the study of microeconomic foundations of macroeconomic relations. Indeed, because of the presence of these shocks the relation between the macro-exogenous variable X_t and whatever macro-endogenous variable Y_t is considered has little chance to be exactly the same as the relation between the corresponding micro-variables x_{it} and y_{it}.

An important particular for explaining microeconomic behaviour and its aggregate implications is what information is available to the individual agent: does he or she observe just ξ_{it} or also each one of its $h+1$ components individually? We shall not argue the point again at this stage, since Section 8.3 did precisely that: the argument there led to explain some degree of price rigidity, but it can clearly be transposed to the many other cases in which information is similarly limited. The last part of this section will present another interesting application of the same idea.

3. In our introduction to the framework proposed by Forni and Lippi we shall now focuss on two issues concerning in particular the stochastic process of the saving rate (see Section 1.8 of Chapter 2). It will appear to the reader that the discussion of these issues is more generally instructive with respect to the aggregation of linear dynamic equations.

Let x_{it} be the earnings of household i in period t and y_{it} be of its consumption. Models of optimum intertemporal consumption explain how household i decides y_{it} from its observation of x_{it} and its knowledge of the stochastic process of future earnings. The typical outcome is existence of a long-run level of consumption proportional ot the long-run level of earnings. In the short-run, however, that level is not reached because of irregular shocks affecting earnings and possibly also consumption. What can be the implications of those models concerning the two-dimensional process of aggregate earnings and consumption?

Motivated by their purpose to analyse how cointegration is affected by aggregation, Forni and Lippi decide in their 1999 article to neglect the presence of idiosyncratic shocks to the earnings process. This would not substantially restrict the analysis if micro-agents would observe precisely the indiosyncratic shocks without any contamination from common shocks. Our authors want, however, to recognize the multiplicity of common shocks, some being permanent others transitory. Three independent white noise processes $\{u_{1t}\}$, $\{u_{2t}\}$, $\{u_{3t}\}$ suffice for their purpose, so that they consider the following equation ruling the earnings process:

$$(1-L)x_{it} = \alpha_i u_{1t} + \beta_i u_{2t} + (1-L)b_i u_{3t}. \tag{360}$$

For simplicity the equation is written without the immaterial constant term; α_i, β_i and b_i are numerical coefficients and L the lag operator. The shocks u_{1t} and u_{2t} are permanent since they affect $x_{it} - x_{i,t-1}$ in period t with no later reversion. In contrast the shock u_{3t} is transitory since after having increased by $b_i u_{3t}$ in period t, the earnings x_i will decrease by $b_i u_{3t}$ in period $t+1$. Behaving in conformity with an optimal intertemporal consumption model and observing the values taken by the shocks, household i decides its consumption y_{it}, which is then ruled by:

$$(1-L)y_{it} = \delta_i \alpha_i u_{1t} + \delta_i \beta_i u_{2t} + (1-L)a_i u_{3t}. \tag{361}$$

The two permanent common shocks, which combine into a permanent shock $\alpha_i u_{1t} + \beta_i u_{2t}$ to earnings, imply a permanent shift in consumption by the amount $\delta_i(\alpha_i u_{1t} + \beta_i u_{2t})$, δ_i being a coefficient whose value is appropriate to the case of household i. Similarly, the transitory common shock u_{3t} implies a transitory shift in consumption by the amount $a_i u_{3t}$.

Let us now consider the two-dimensional process of (x_{it}, y_{it}). It is integrated of order 1 since x_{it} and y_{it} follow processes which are difference-stationary but not stationary (see Section 7.4). However, the process is cointegrated because (360) and (361) imply:

$$y_{it} - \delta_i x_{it} = (a_i - \delta_i b_i) u_{3t}. \tag{362}$$

Aggregating the x_{it} and y_{it} into their respective sums X_t and Y_t, we find that $(1 - L)X_t$ and $(1 - L)Y_t$ obey equations similar to (360) and (361), except that the six coefficients of u_{1t}, u_{2t} and $(1 - L)u_{3t}$ are replaced by their respective sums. The two-dimensional process of (X_t, Y_t) is integrated of order 1. But, given heterogeneity of the households and of the two permanent common shocks, cointegration is quite unlikely. Indeed, cointegration would mean existence of a number c such that $Y_t - cX_t$ be stationary. This in turn would require:

$$\sum_i (\delta_i - c)\alpha_i = 0, \qquad \sum_i (\delta_i - c)\beta_i = 0, \tag{363}$$

since u_{1t} and u_{2t} are ruled by independent random processes. There are two possibilities for existence of a number c fulfilling the two equations (363): either all δ_i are equal to the same number δ and then (363) holds with $c = \delta$, or there is a number τ such that, for all i, $\beta_i = \tau\alpha_i$ and $c = \sum_i \delta_i \alpha_i / \sum_i \alpha_i$ fulfils not only the first equation (363) but also the second. Let us see why neither of these two possibilities is likely, given heterogeneities actually faced.

The first possibility would mean that all households would react in the same way to a given permanent shock to their earnings, no matter how risk averse they may be, how many children they have or how advanced they are in their life cycle. This is not likely. The second possibility would mean that, in the permanent earnings shocks $\alpha_i u_{1t} + \beta_i u_{2t}$ faced by various households, the two common shocks would appear with the same respective weights. For instance, if the population was made only of farmers and wage-earners, u_{1t} could be a measure of technologically induced shifts in the real value of agricultural output and u_{2t} a measure of shifts in the real wage rate: the relative weight of u_{1t} would certainly be high for farmers and low for wage-earners. (Clearly many more dimensions of heterogeneity must be recognized than those just mentioned.)

Intuitively we understand that aggregation blurs long-run cointegrating relations which microeconomic modelling finds to be neat and sharp. How far can such a blurring go? That seems to be first and foremost an empirical question.

4. In the introduction to their book, Forni and Lippi suggest that recognizing aggregation effects at times plays a positive role in macroeconomics. Many formal results about these effects appear negative since they invalidate simple transpositions of the properties established about the rational behaviour of the individual household or firm. But other results explain why and how aggregate data, which also disprove the simple transpositions, can be compatible with rational microeconomic behaviour. Such results are truly positive contributions to macroeconomic theory building.

The authors take as examples explanations of aggregate empirical evidence which seems, at first sight, to disprove the "permanent income hypothesis" as it is derived from the study of the stochastic process of consumption of an individual household (see Section 1.8 in our Chapter 2). This is the subject of their Chapter 13 which runs over thirty pages. Here, we shall limit attention to just one so-called puzzle, namely the "excess smoothness" which is reported to be found in aggregate time series of consumption, and to just one simple explanation borrowed from the Forni–Lippi 1999 article.

A. Deaton showed that, if the earnings process exhibits high persistence, which is indeed observed in aggregate data, then according to the permanent income hypothesis consumption should be more volatile than earnings. But such is actually not the case. The most elementary argument proposed by Forni and Lippi takes advantage of results coming from microeconomic data and interpret them within an extremely simple case of their more general models.

Let the earnings process obey:

$$x_{it} - x_{i,t-1} = u_t + a u_{t-1} + \eta_{it} + b \eta_{i,t-1}, \tag{364}$$

where u_t and η_{it} are two independent white noises with respective variances σ_u^2 and σ_η^2 (the same for all i). With respect to (360) we note that an idiosyncratic shock is present, but that no other heterogeneity appears: in particular there is just one common shock, the coefficients a and b are the same for all households, all idiosynchratic shocks have the same variance, they are moreover assumed mutually independent. The shock processes are first-order moving averages.

Given the form of the right-hand member of (364) there is a number d and a white noise ε_{it} with variance σ_ε^2 such that (364) may also be written as:

$$x_{it} - x_{i,t-1} = \varepsilon_{it} + d \varepsilon_{i,t-1}. \tag{365}$$

(In the theory of stochastic processes this is known as "the Wold representation" of $x_{it} - x_{i,t-1}$ if the absolute value of d is smaller than 1.) The numbers d and σ_ε^2 have to satisfy the two following equations, which derive directly from equality of the two right-hand members of (365) and (364):

$$\begin{aligned} &\left(1+d^2\right)\sigma_\varepsilon^2 = \left(1+a^2\right)\sigma_u^2 + \left(1+b^2\right)\sigma_\eta^2, \\ &d\sigma_\varepsilon^2 = a\sigma_u^2 + b\sigma_\eta^2. \end{aligned} \tag{366}$$

These two equations actually determine the appropriate numbers d and σ_ε^2. Indeed, elimination of σ_ε^2 leads to:

$$d^2 - Ad + 1 = 0 \tag{367}$$

with

$$A = \frac{(1+a^2)\sigma_u^2 + (1+b^2)\sigma_\eta^2}{a\sigma_u^2 + b\sigma_\eta^2}. \tag{368}$$

Equation (367) has two reciprocal roots. The appropriate root has modulus smaller than 1. The second of Equations (366) gives the appropriate value of σ_ε^2. Comparing (365) to (364) we moreover see that ε_{it} obeys:

$$\varepsilon_{it} = \frac{(1+aL)u_t + (1+bL)\eta_{it}}{1+dL}. \tag{369}$$

Let us assume further that households observe only the variations in their earnings and are so unable to identify separately what comes respectively from the common shock u_t and from the idiosyncratic shock η_{it}. Then, applying the model that led us to Equation (86) in Chapter 2 and taking the intertemporal budget constraint into account as well as (365), we may interpret the permanent income hypothesis as implying:

$$c_{it} - c_{i,t-1} = (1+\beta d)\varepsilon_{it} \tag{370}$$

in which β is the market discount factor, assumed to be constant over time. Taking (369) into account we find:

$$c_{it} - c_{i,t-1} = \frac{1+\beta d}{1+dL}\left[(1+aL)u_t + (1+bL)\eta_{it}\right]. \tag{371}$$

Equations (364) and (371) characterize the micro-model. Taking arithmetic means and applying the law of large numbers we derive the macro-model linking the process of $\bar{c}_t$ to that of $\bar{x}_t$, namely:

$$\bar{x}_t - \bar{x}_{t-1} = u_t + au_{t-1}, \qquad \bar{c}_t - \bar{c}_{t-1} = \frac{1+\beta d}{1+dL}(u_t + au_{t-1}), \tag{372}$$

which we may also write as an autoregression:

$$\bar{c}_t - \bar{c}_{t-1} = -d(\bar{c}_{t-1} - \bar{c}_{t-2}) + (1+\beta d)(\bar{x}_t - \bar{x}_{t-1}). \tag{373}$$

Excess smoothness means that the series of changes in consumption, $\bar{c}_t - \bar{c}_{t-1}$, is more smooth than that of changes in earnings, $\bar{x}_t - \bar{x}_{t-1}$. This occurs when d is negative, i.e., when the sum A of the two roots of Equation (367) is negative. At this point Forni and Lippi call on microeconomic empirical evidence obtained by J.-S. Pischke[209], precisely for his study of the impact of imperfect information on the dynamic properties of the aggregate consumption process (for Pischke as well as for Forni and Lippi, the imperfect information of households concerns the origin of changes experienced in earnings). From a US panel of household monthly data, Pischke derived estimates of the dynamic properties of the quarterly series of individual earnings in relation to the quarterly series of aggregate earnings. Translated in our notation these estimates mean that σ_η^2 is much larger[210] than σ_u^2, that b is negative and that the denominator in (368) is also negative (negative autocorrelation prevails in individual earning changes). It then follows that A, hence d, are negative, precisely the condition for excess smoothness.

To conclude: aggregation effects explain the appearance of excess smoothness. However, Pischke finds less excess smoothness than exists in the aggregate series. This means that we do not reach at this point a complete explanation of the phenomenon. But the literature partially surveyed here in Chapter 2 contains other explanations, which complement the one just given.

[209] J.-S. Pischke, Individual income, incomplete information, and aggregate consumption, *Econometrica* (July 1995).

[210] As pointed out, by Pischke, the fact that "is much larger than" implies that most individual households have only a negligible incentive to take account of aggregate information about earnings when deciding about their own savings: neglecting the public information about aggregates, which moreover arrive with a lag, is hardly irrational.

This Page Intentionally Left Blank

Chapter 9

Business Fluctuations and Stabilization Policies

Observing how business conditions evolve reveals substantial but somewhat irregular fluctuations concerning production, employment, prices and most other macroeconomic variables. Movements of these various magnitudes are mutually linked and, to a large extent, due to common causes. The two preceding chapters indeed concerned the phenomenon and its origins. They bore more particularly on changes in output and in the rate of inflation.

But those two chapters did not exhaustively deal with the full phenomenon. They did not intend to encompass all relevant variables. They neglected the cyclical appearance of business fluctuations, an aspect which was on the contrary often stressed and the focus of attention since long in the past. In order to now complete our study we must approach the phenomenon from a different side.

In the same way as in Chapters 7 and 8 macroeconomic theory will have three main concerns: understanding the facts, developing tools for their current analysis, judging policy interventions. In the same way, our account of the theory will have to pay attention first to what was considered as common wisdom in the 1960s, when business fluctuations were seen as due to the spontaneous dynamics of aggregate demand, occasionally hitting an upper limit imposed by supply, and when macroeconometric models were developed in order to serve for business diagnosis and policy analysis. But we shall have also to consider the outcome of many developments in macroecomic research during the last thirty years, which led to much more circumspect and eclectic conceptions, both in academic circles and where macroeconomic theory must be applied.

Accordingly this chapter will have four parts, respectively entitled: 1. The Keynesian vision of business cycles; 2. Business cycle facts: new interpretations; 3. Structural macroeconometric models; 4. Alternative econometric approaches to diagnosis and policies.

1 The Keynesian Vision of Business Cycles

Many theories had been thought up in the past in order to explain economic fluctuations. A systematic survey of them, first published in French in 1936, still gives an excellent land-mark on the state of economic ideas at the time[1]. We shall consider it later in Section 2.10. Indeed, diffusion of Keynesian theory shortly after brought a radical simplification, which is still meaningful and will concern us in this part of the chapter. At the very least, the Keynesian vision[2] will give us a reference to which more complex or alternative business cycle conceptions will be compared. Let us then consider what was standard teaching in the 1960s, starting from its preliminaries.

The actual evolution in an economy is obviously not independent of changes which affect its environment, whether this be physical, technological, political or social, or for any particular country, the foreign economic environment. It is therefore conceivable, *a priori*, that cyclical fluctuations could be entirely explained by cyclical changes in this environment.

Indeed, alternating years of abundance and scarcity have for a long time been blamed on meteorological causes, as late as the 19th century even in industrial countries. The ebb and flow of the great contagious diseases played an obvious role until the 17th century. More recently, irregularity in the discoveries of either mineral resources, or in industrial technological progress has notice-

[1] G. Haberler, *Prosperity and Depression* (League of Nations, Geneva, 1937); later editions at Harvard University Press, 6th edition in 1964.

[2] Here, in the same spirit as in Chapter 7, we speak of the Keynesian vision when referring to ideas of Keynesian economists of the 1950s and 1960s.

ably affected the economic climate. Everywhere armed conflicts had effects on economies. In some countries, it could be said that the alternation of political parties, right and left, in power has been reflected in changes in employment and inflation[3].

If such considerations were sufficient to explain cyclical fluctuations, there would be no need for a specific economic theory. Fluctuations in production, employment and prices would simply reflect the changes in the variables that economists properly treat as exogenous. The theory of short-term equilibrium, the subject of Chapter 7, would explain how changes in exogenous variables were reflected in changes in such endogenous variables as output, employment and prices.

But nowadays, we believe that cyclical fluctuations are not a simple reflection of changes in the environment. The paths and patterns followed by variables also depend on economic factors, in a way that has to be understood. We may even claim that, at times, such were at the origin of the movements observed.

With a view to clarifying the economic aspect of the phenomenon, it is legitimate to first rule out changes in the environment and to concentrate on what the evolution of the economy would be without such changes, and also without any sort of policy intervention. This is how we shall proceed in this part of the chapter. We shall rely on fairly simple theoretical models inspired by the Keynesian approach. We shall also make reference to stylized facts borrowed from observed evolutions, but empirical justification of these facts will come later in the chapter.

We shall start by giving a verbal account of the essential elements of the phenomenon, considering in particular the various phases of the economic cycle. Then we shall examine, one by one, some of the factors which should feature in a complete theory of business cycles: the formation of demand, the limitation of productive capacities, the distribution of incomes, changes in prices and interest rates, liquidities held by firms and households. We shall thus prepare the way for further developments.

1.1. First elements on economic fluctuations

Changes in aggregate demand make the most obvious feature of the business cycle. In periods of expansion, demand increases strongly, stimulating rapid increases in production. Most often lack of equipment or available labour begin

[3] See, for example, D. Mac Rae, A political model of the business cycle, *Journal of Political Economy* (April 1977).

to appear after some time in particular industries or regions. The resulting excessive pressure of demand on some markets generates price increases and this tones down demand. Somehow the switch back, which initiates a relaxation of the pressure on productive capacities, does not stop there: the reduction in demand accelerates with the onset of the pause, which then turns into a more or less serious depression. Finally, recovery follows when demand becomes firmer.

It has been observed for a long time that, among the different components of aggregate demand, investment and accumulation of inventories fluctuate the most[4]. The changes in these two components therefore play an essential role and should feature in any explanation of cycles. The first reason for those fluctuations in investment and inventories comes from the fact that, in the short run, the need for equipment and inventory holding is almost proportional to production. So, firms have to invest and stock a lot when output increases rapidly; they on the contrary feel little need to invest and they may even try to run stocks down when demand falls. Once started, any fluctuation in output brings about a change which is relatively much greater in investment and accumulation of inventories, and this accelerates the initial movement. We shall devote the following section to a precise study of this.

As long as we shall focus only on the need to invest and to build inventories we shall be led into neglecting changes in costs, in the financial resources available to the firms which invest and even in the physical capacities of the firms which supply capital goods. When needs for such goods become too intense and general, we may expect their cost to rise, the investing firms' available finance to often be exhausted, and demand to be faced with bottlenecks in the production of equipments and storable goods. As we shall see, these problems raise complex issues. Nevertheless, we understand from the outset that there are limits which prevent the acceleration phenomenon from continuing indefinitely. We shall examine the origin and effects of these limits in Section 1.3.

It would moreover be a mistake to neglect changes in households' consumption and saving. On the one hand, we understood in Chapter 7 that the simple multiplier amplifies movements which the accelerator generates in productions and incomes; we shall take this into account already in Section 1.2. On the other hand, cyclical variations induce changes in income distribution; as various groups of households have different propensities to consume, induced changes in aggregate demand may also occur on that score.

These have two main features. In the first place, the distribution, between wages and profits, of incomes from production changes during the cycle. The

[4] Exports also fluctuate a great deal. This is quite important, in practice, in small economies which are most dependent on foreign trade.

share of profits tends to increase at the end of a depression and during a recovery. The share of wages tends, on the contrary, to increase at the end of a boom and the beginning of a depression; this movement increases the pressure on demand, which first worsens the tension but later compensates somewhat for the decrease in investment and stocking. We shall examine these aspects more precisely in Section 1.5, after developing a method for doing this in Section 1.4.

In the second place, the rules governing transfer payments, through either taxes or benefits paid to households, compensate partially, but fairly significantly, for the changes in incomes directly resulting from production. In this way, cyclical movements are naturally dampened. This is the role of the "automatic stabilizers" which will be briefly considered at the beginning of Section 1.5.

Business cycles matter mainly because of the changes they produce in the level of economic activity; but they also show up in oscillations in the rate of inflation and in some relative prices. These movements are interesting in themselves. In addition, they react on the cycle through the effect they have on the purchasing power of wealth, on inventory decisions or on the respective costs of labour and capital. We shall see this in Section 1.6.

Finally, we shall later have to consider those monetary phenomena which in the past have often been given the main responsibility for economic cycles because of their effect on the relative abundance or scarcity of credit. Section 1.7 will not go into such a study but just give a quick look at the demand for money.

This rapid review of the main features of cycles shows the diversity of the factors which are involved in the explanation of cyclical fluctuations. Besides, we may expect the relative weight of each of these factors to vary in terms of very many circumstances and so the "cycles" to vary very much in their magnitude, length and other features. Our object, in this part of the chapter, is not to construct a complete theoretical prototype which would integrate and put exactly at their respective places all the factors to be considered. It is rather to provide a first analysis to be further developed in the following parts.

1.2. Interplay of the multiplier and the accelerator

1.2.1. *A model of the phenomenon*

Investment and the accumulation of inventories amplify exogenous fluctuations in demand and production. We can see this by looking at very simple and almost mechanical equations, before we introduce complications in the following sections[5].

Let us go back to the theory of the multiplier presented in Chapter 7. There it was expressed in a differential form, Equation (37), for example, leading to:

$$\mathrm{d}y = \frac{p}{a_y}\,\mathrm{d}G \tag{1}$$

if, of the exogenous variables, only public consumption G changes, and neither prices nor the interest rate are affected.

Here, we are going to keep *the assumption that changes in demand do not influence prices nor the rate of interest*, but only quantities. That is, we shall not pay attention for the time being to complications appearing when the pressure of demand on productive capacities becomes acute.

The coefficient a_y/p, the inverse of the multiplier, was defined by:

$$\frac{a_y}{p} = 1 - (1 - \varepsilon - \tau) f'_Y - g'_y, \tag{2}$$

where ε and τ were the ratios, assumed to be fixed, of firms' saving (or self financing) and of taxes to output. So in our study of the short-term equilibrium, the marginal propensity to consume f'_Y and to invest g'_y came from "static" models in which behaviours were assumed to react to simultaneous changes in some variables.

Now, when we studied the econometric laws of consumption and investment, we saw that "dynamic" effects played an important part, one which we can no longer ignore, the lags in adaptation being an essential feature in business cycles. Chapter 3 thus highlighted Equation (255) as a simple expression of the consumption function. We can write it here as:

$$C_t = \alpha(1 - \gamma)\frac{Y_t}{p_t} + \gamma C_{t-1} + c. \tag{3}$$

[5] Models expressing the interplay of the multiplier and the accelerator became classic already long ago. They may be traced back to a number of economists listed in Chapter 3, Part C of G. Haberler, *Prosperity and Depression*, op. cit. Here we may refer in particular to J. Clark, A non-monetary cause of fluctuation in employment, *Economic Journal* (September 1914). The mathematical treatment was discussed by P. Samuelson, Interactions between the multiplier analysis and the principle of acceleration, *Review of Economic Statistics* (1939); reproduced in American Economic Association (ed.), *Readings in Business Cycle Theory* (Irwin, Homewood, IL, 1951).

Similarly Chapter 4 exhibited Equation (262) for the acceleration to which investment is subject. We can transpose it here as:

$$I_t = \sum_{\tau=0}^{h} \alpha_\tau b(y_{t-\tau} - y_{t-\tau-1}) + \delta b y_{t-1}. \tag{4}$$

We saw, moreover, also in Chapter 4 that a similar law could be used for the accumulation of inventories.

If we stick to the same idea of the model used for the theory of the multiplier, we can easily incorporate into it the notion of acceleration and that of lags in consumption with respect to income. To obtain a simple expression, we may for the moment ignore the difference between real income and production (Section 1.5 will study more precisely how income distribution intervenes). In addition, we keep to purely linear equations, which is sufficient at this stage (see Section 1.3); as in Chapter 8, we use the symbolic operator L defined by $Lx_t = x_{t-1}$.

Finally, A_t will denote "autonomous demand", that is, the part of aggregate demand which, in the immediate economic context, may be considered as being independent of the level of output and of its changes: this may cover government consumption and a good part of investment or even of consumption.

Thanks to these different conventions we can replace Equation (1), expressing the theory of the multiplier, by equation:

$$\left[1 - C(L) - I(L)\right] y_t = A_t, \tag{5}$$

where $C(L)$ and $I(L)$ are two functions of the symbolic operator L, so that $C(L)y_t$ and $I(L)y_t$ are the non exogenous parts of consumption C_t and of investment I_t, respectively. For example with Equations (3) and (4) we have:

$$C(L) = \frac{\alpha(1-\gamma)}{1-\gamma L}, \tag{6}$$

$$I(L) = \sum_{\tau=0}^{h} \alpha_\tau b L^\tau (1-L) + \delta \cdot bL. \tag{7}$$

We now have to study how, from Equation (5), the changes affecting exogenous demand A_t are transformed into fluctuations in output y_t. This study has a formal, mathematical, side which it is convenient to deal with before considering economic implications. For so doing we shall keep, here, to the case where $1 - C(L) - I(L)$ is a rational function of L, which is obviously not restrictive in practice.

1.2.2. *Mathematical digression*

We have already approached this mathematical question in Chapter 8 when we considered the stability of the inflationary process (see, for example, Section 4.1 of Chapter 8). We are now going to examine, briefly but in general terms, the dynamic link between an exogenous variable z_t and an endogenous variable x_t.

$$\frac{Q(L)}{R(L)} x_t = z_t, \tag{8}$$

where $Q(L)$ and $R(L)$ are two polynomials.

Studying this link will be made easier by introduction of the intermediate variable:

$$w_t = R(L) z_t. \tag{9}$$

The variable in question is a linear function of the successive values of z_t, for example, $z_t, z_{t-1}, \ldots, z_{t-T}$. In the case where $R(L)$ would have all its coefficients positive, w_t would thus be, up to a multiplicative constant, a weighted average of the successive values of the z_t; the sequence of the w_t would thus correspond to a smoothing of the sequence of the z_t. In the case where $R(L)$ would be equal to $1 - \gamma L = (1 - \gamma) + \gamma(1 - L)$ in which γ would be a number between 0 and 1, the variable w_t would be a weighted average of the level z_t and of the increase $z_t - z_{t-1}$. At all events, when the polynomial $R(L)$ is low order, the variable w_t is influenced by a small number of successive values of z_t and not by the distant past history of this variable.

Elucidating how x_t depends on w_t through the relationship:

$$Q(L) x_t = w_t \tag{10}$$

is a little more delicate. We can start from a decomposition of polynomial $Q(L)$ into its elementary factors:

$$Q(L) = q_0(1 - a_1 L) \ldots (1 - a_h L), \tag{11}$$

where $a_1, a_2, \ldots, a_h$ are the inverses of the roots of $Q(L)$. In particular, we can easily deal with the case where $h = 1$; for the recursive Equation (10) then has the solution:

$$x_t = a_1^t x_0 + \sum_{\tau=0}^{t-1} a_1^\tau w_{t-\tau}, \quad t = 1, 2, \ldots .$$

We can see that if a_1 is not less than 1 in absolute value, the initial value x_0 will have an influence which will not fade away as t increases. So, stability requires $|a_1| < 1$. More generally, it requires that moduli of all the roots of $Q(L)$ be greater than 1 (that the $|a_j|$ be lower than 1).

To understand the essence of the problem, we shall write (10) in the form:

$$b_0 x_t + b_1 x_{t-1} + \cdots + b_h x_{t-h} = w_t \tag{12}$$

and assume that b_0 is different from zero, which we can always achieve, if necessary by changing the definition of w_t and of h. This is not the place to go through the complete theory of this type of finite linear difference equation with constant coefficients, a theory which is, moreover, analogous to that, still better known, of linear differential equations with constant coefficients. Let us have another look at the main elements.

(i) We call the "general solution" of Equation (12) an equation which expresses x_t in terms of the $w_{t-\tau}$ and which applies no matter what the initial values of x may be. As opposed to that, a "particular solution" is a sequence of values of the x_t, which satisfies (12) for given initial values.

(ii) The general solution of (12) can be expressed as the sum of any particular solution to this equation and of the general solution of the homogeneous equation corresponding to (12):

$$b_0 x_t + b_1 x_{t-1} + \cdots + b_h x_{t-h} = 0 \tag{13}$$

(this is also said to be "the equation with no right-hand member").

(iii) The general solution of this last equation is any linear combination of h particular solutions which are linearly independent of each other.

(iv) If the "characteristic equation" in u:

$$b_0 u^h + b_1 u^{h-1} + \cdots + b_h = 0 \tag{14}$$

has the solution a_j, real or complex, then

$$x_t = a_j^t \tag{15}$$

obviously defines a particular solution of the equation with no right-hand member (note that any number a_j appearing in the factorization (11) above is a solution of (14)). If, in addition, the h roots of (14) are distinct, the h solutions (15) are linearly independent. If a_j is a multiple root of the order of n, then

$$x_t = t^m a_j^t \tag{16}$$

defines a particular solution of (13) for any positive integer $m < n$; such supplementary solutions are independent of each other and independent of solutions (15). In any event, since b_0 differs from zero, we can so find a system of h particular solutions of the equation with no right-hand member, solutions which are mutually linearly independent.

To study the time-profiles of the solutions of Equation (12) it is useful to turn attention to the roots of Equation (14) and to solutions such as (15) and (16). In the case where a_j is complex, these solutions do not seem to be very interesting; but the difficulty can be eliminated by simultaneously considering a_j and its conjugate (also a solution of (14) since this equation has only real coefficients). Indeed we may write:

$$a_j = \rho_j[\cos\omega_j + i\sin\omega_j],$$

where ρ_j and ω_j are the modulus and the argument of a_j, whereas i is the complex number with square equal to -1. The conjugate root can be written:

$$a_{j+1} = \rho_j[\cos\omega_j - i\sin\omega_j].$$

Through linear combinations of a_j^t and a_{j+1}^t we can define two particular solutions which are real and linearly independent, namely:

$$x_t = \rho_j^t \cos\omega_j t, \quad x_t = \rho_j^t \sin\omega_j t. \tag{17}$$

Solutions of this type, which correspond to complex roots, thus have damped oscillations if $\rho_j < 1$, explosive oscillations if $\rho_j > 1$. We can obviously operate in the same way with solutions of type (16).

Be that as it may, considering the particular solutions of (13) leads us to conclude that the general solution of this equation with no right-hand member tends to zero if all the roots of Equation (14) have moduli less than 1. In other cases the general solution features components which either diverge, or remain constant ($a_j = 1$), or oscillate with a constant amplitude (a complex root with modulus equal to 1).

Coming back to model (5) we shall see how these general principles can be applied to the case to which we were naturally led, right from the start. We know that the time path of x_t will depend very strictly, even from a qualitative point of view, on Equation (14), that is, on the specification of the lags and the coefficients which feature in $Q(L)$. We shall therefore have to be precise in formulating our model in order to arrive at clear conclusions[6].

1.2.3. *A specification*

The remark just made suggests that preliminary econometric research might be advisable before we study the dynamic properties of an economy where only the multiplier and the acceleration phenomenon would matter. Econometric research would then aim at estimating the values of α, γ, h, b, δ and the α_τ in the behavioural equations (3) and (4). But there would be no great point in trying to reproduce exactly the econometric equations so obtained since, in actual economies, other factors than those studied in the present section come into play.

This is why the study of econometric models applying to such or such a country, will be postponed until Part 3. For the moment we shall examine just an example, whose simplicity makes both calculations and understanding easier, but which nevertheless retains, for its behavioural equations, the main characteristics of the econometric equations otherwise obtained. The qualitative properties of the resulting time paths will so merit some attention.

Since we have to juxtapose the equations relating, on the one hand to consumption, and on the other to investment, for example as specified in (3) and (4), we have to take into account the fact that the second includes lags which are much longer than the first: consumption adapts to income much more quickly than the stock of capital adapts to the level of production. In order to keep to a very simple model, we shall put together Equation (6) for consumption, and the following particular specification of (7) for investment[7]:

$$I(L) = \frac{(1-\gamma)bL(1-L)}{1-\gamma L} + \delta \cdot bL \tag{18}$$

[6] Indeed, no mathematical result specifies general conditions for the b_τ that could be easily applied to the phenomenon considered in this section and would unambiguously identify the type of time path which the general solution of Equation (13) would follow. Such conditions, which are applicable to other cases, have been given by R. Sato, for example, in: A further note on a difference equation occurring in growth theory, *Journal of Economic Theory* (March 1970).

[7] This special form specifies quite strictly the lags in the investment equation, in relation to those in the consumption equation. If the basic period is a year and if $b = 1/4$, for example, the average lag with which I_t reacts to $y_t - y_{t-1}$ is about a year and a half, as in Equation (264) of Chapter 4. But this action is much less staggered over time. However, the specification chosen is important in what follows.

(in Equation (7), the coefficient α_0 then is zero whereas the other α_τ are equal to $(1-\gamma)\gamma^{\tau-1}$). It is obviously artificial to use the same number γ in Equations (6) and (18), but it greatly facilitates the following discussion.

To sum up, we assume that consumption reacts immediately to changes in income. A first lag of one period occurs on the other hand before investment adapts at all, whether this concerns the replacement or the expansion of productive capacity. Even though it is somewhat abridged, this assumption will suffice for a theoretical exploration of the interplay between the multiplier and the accelerator.

We can now write Equation (5) in the form of Equation (8) in order to simply express the dynamics which result from the formation of aggregate demand:

$$\frac{1-\alpha(1-\gamma)-[\gamma+b(1+\delta-\gamma)]L+b(1-\gamma+\gamma\delta)L^2}{1-\gamma L}y_t = A_t. \tag{19}$$

Studying this difference equation amounts, as we have seen, to studying the solutions of Equation (14) in u, which here, takes the expression:

$$\begin{aligned} P(u) = {} & (1-\alpha+\alpha\gamma)u^2 - \left[\gamma+b(1-\gamma+\delta)\right]u \\ & + b(1-\gamma+\gamma\delta) = 0. \end{aligned} \tag{20}$$

As α and γ are positive and smaller than 1, we can check immediately the following inequalities:

$$P(-\infty) > 0, \qquad P(0) > 0, \qquad P(\infty) > 0.$$

In addition, $b\delta$ has to be small, so that we can also assume:

$$P(1) = (1-\gamma)(1-\alpha-b\delta) > 0.$$

The minimum of $P(u)$ is attained for the value u_m, which is obviously positive, given by:

$$2(1-\alpha+\alpha\gamma)u_m = \gamma + b(1-\gamma+\delta). \tag{21}$$

We can now distinguish the following cases:

(i) If Equation (20) has two complex conjugate roots, that is if $P(u_m) > 0$, their common modulus ρ is given by:

$$\rho^2 = \frac{b(1-\gamma+\gamma\delta)}{1-\alpha+\alpha\gamma}. \tag{22}$$

Depending upon whether ρ is smaller or greater than 1, the general solution of the equation corresponding to (19) with no right-hand member has oscillations which are either damped or explosive.

(ii) If Equation (20) has two real roots, which may be identical, that is if $P(u_m) \leqslant 0$, these two roots, which are necessarily positive, will both be less or greater than 1 depending upon whether u_m is less or greater than 1.

To be more specific, we should start from an assertion stating what might be the likely values of the coefficients α, γ, δ and b. But even an extensive discussion would not be of much help. Of all coefficients, b is without doubt the least well determined by econometric studies. Hence we are going to fix the values of the three other coefficients, trying only to give roughly acceptable values to these parameters.

Considering the case where the unit time period is a year, we shall concentrate on the following values:

$$\alpha = 2/3, \qquad \gamma = 1/4, \qquad \delta = 1/10. \tag{23}$$

We can thus calculate:

$$u_m = \frac{5+17b}{20}, \qquad \rho^2 = \frac{31}{20}b.$$

We can also see that Equation (20) has real roots if b is greater than a value β_2 approximately equal to 1.2. In sum, we end in distinguishing three cases:

- If $b < \beta_1 = 20/31$, the time path of aggregate demand (and production) exhibits damped oscillations.
- If $\beta_1 \leqslant b < \beta_2$, the time path still oscillates but more so from one cycle to the next (except if $b = \beta_1$). We say that it exhibits amplified oscillations.
- If $b \geqslant \beta_2$, the time path is monotone, but divergent. We say that it is "explosive".

1.2.4. *Relevance of the reference time path*

The above analysis is obviously subject to two reservations which we must keep in mind.

In the first place, we gave only the characteristics of the time path of the general solution of the equation with a right-hand member equal to zero. The actual development of y_t will depend on the time path followed by the actual right-hand member, here $A_t - \gamma A_{t-1}$, a weighted average of autonomous demand and the change in it. Nevertheless, the discussion above gave us an interesting insight.

Let us assume, for instance, that A_t was constant in the past ($A_t = A_0$ for all $t \leqslant 0$) and that y_t also remained constant at a value y_0; Equation (19) thus requires:

$$y_0 = \frac{A_0}{1 - \alpha - b\delta} \tag{24}$$

this is the effect of the (long-run) multiplier. Assume that, from date $t = 1$, A_t is again constant and equal to a value A_1 different from A_0. How does y_t evolve from $t = 1$ onwards?

To find out, we just have to calculate, with the help of Equation (19):

$$y_1 = \frac{A_0}{1 - \alpha - b\delta} + \frac{A_1 - A_0}{1 - \alpha + \alpha\gamma}.$$

We see then that for $t \geqslant 2$, Equation (19) has a constant right-hand member equal to $(1 - \gamma)A_1$. Thus, a particular solution of this equation is:

$$y_t^a = \frac{A_1}{1 - \alpha - b\delta}.$$

But, for the actual time path of y_t, we have to add to this expression a solution y_t^b of the equation with no right-hand member and this must be chosen in such a way that, for $t = 0$ and $t = 1$, the sum $y_t^a + y_t^b$ takes the values y_0 and y_1 determined above. We can calculate with no difficulty:

$$y_0^b = \frac{-(A_1 - A_0)}{1 - \alpha - b\delta}, \qquad y_1^b = \frac{-(\alpha\gamma + b\delta)(A_1 - A_0)}{(1 - \alpha - b\delta)(1 - \alpha + \alpha\gamma)}.$$

Since y_t^a is constant, the time path of y_t will be analogous to that of y_t^b, whose evolution was studied above, except for the two first values. More precisely, y_t will tend to the "equilibrium value" $A_1/(1 - \alpha - b\delta)$ after dampened oscillations if the parameter values given by (23) apply and if b is smaller than β_1. On the other hand, if b is greater than β_1, production y_t will not converge to this equilibrium value; the oscillations will be amplified from one cycle to the next, or even the gap between production and its equilibrium value will be made wider and wider, all along.

In reality, A_t is highly likely to follow a fairly irregular path. For studies similar to ours here, it has even often been suggested that autonomous demand could be viewed as ruled by a stochastic process. In any case, we intuitively understand that, since production is given by a linear equation, its path over time is obtained by adding terms similar to those we just described. In the end, it is still the general solution of the equation with no right-hand member which dictates the qualitative characteristics of the relevant particular solution of Equation (19)[8].

In the second place, the paths which this simple model leads us to are meant to be followed only in periods during which the model provides a valid first approximation, that is, when changes in aggregate demand are directly reflected in changes in output: the conditions required for the theory of the multiplier to hold in a static perspective apply here, even though the perspective adopted is now dynamic.

This remark implies, in particular, that these solutions could not be prolonged indefinitely when they exhibit amplified oscillations or an upward explosive evolution. At one moment or another, demand would come up against the limits of productive capacities and the change in production would then have to obey other equations than those used here. This type of complication will be the subject of the following section; so we are not going to discuss it further now.

1.2.5. *An oscillatory reference path*

Even though we have chosen a model which is not only simplifying but also special, can we still use it to get a first idea of the dynamics of business cycles in real economies? Certainly, the question will not be properly answered before comparison can be made with fairly complete econometric models. But a least, just to start us thinking, we can ask what value b should have in Equation (19), and then which of the three possible cases, depending upon the position of b with respect to β_1 and β_2, would be most realistic.

When we constructed the theory of the accelerator which led us to Equation (7), we started from the idea that the desired capital K was proportional to production y; the number b was then the desired value of the capital–output ratio. Evaluations given for this coefficient at the national level vary from 2 to 4, according to the definition chosen (what is recorded under the label of

[8] This assertion can obviously be made rigorous, but we are not trying here to go deeper into the mathematical theory, which would lead us too far from our main point.

"national capital" is obviously very important). If we stick to these aggregate evaluations, we have to conclude that, since b is clearly higher than β_2, the reference time path ought to be "explosive".

But a little reflection shows that the value we give to b in Equation (18) has to be much lower than that of the national ratio. Indeed, this equation is used here with a value of γ equal to $1/4$ and a unit time period of one year. Thus, the ratio denoted as b concerns only those equipments which adjust fairly rapidly to the level of production, with a maximum lag of three or four years, say. The basic infrastructure is obviously not taken into account.

So, what should be the value given to b in our model? We propose to consider the sum of the coefficients appearing in available econometric estimations in the terms relating to the accelerator. Thus, according to Equation (264) of Chapter 4, which applied to industries with capital–output ratios varying between 1 and 2, the sum in question would be between 0.6 and 1.2. The estimates found in the econometric literature usually give lower values[9].

We are not aiming for great precision, which we could not claim in any case, since we are using a simplifying and special model. Nevertheless, we can still argue that, in so far as the model may provide a useful approximation, b ought to be lower than β_2. Then the reference time path ought to exhibit oscillations, whether damped or amplified.

The results of this rough discussion will then appear most simply if we examine the case where the reference path, is purely sinusoidal[10]. This occurs when $\rho^2 = 1$, thus $b = 20/31$. Accepting for α, γ and δ the values given by (23), we find that the two particular solutions of the form (17) involve an angle ω of $57°23'$, which can be rounded to $\pi/3$. Thus the cycle would have a period of about six years[11].

Figure 1 shows the approximate reference paths of consumption, investment and production (so as to avoid considering the negative values of these variables, we should think of the curves of the figure as tracing the difference in the values from their long-run equilibrium trends). Like production, consump-

[9] Note, also, that the lags are more spread out in Equation (264) of Chapter 4 than in Equation (18) of this chapter, even though the average lag is about the same. This should lead, here, to a preference for a lower value for b than the sum of the coefficients of the acceleration terms which appear in the econometric fit.

[10] That the reference cycle has a sinusoidal form, whether damped or amplified, comes from the fact that we selected a second degree equation. In the general case where Equation (14) has a higher degree, a reference cycle would combine several sine curves superimposed and would, therefore, have a different profile.

[11] This happens to be not too bad as an order of magnitude for the period of post-war European business fluctuations. In France, for instance, clear peaks occurred in 1951, 1957, 1963, 1969, 1973, 1979, 1989 and 1994.

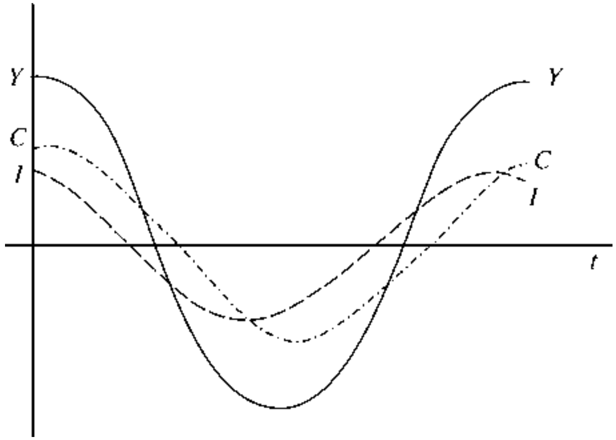

Fig. 1.

tion and investment have sinusoidal paths, with consumption having a lag of about a quarter and investment a slightly longer lead.

We examined the lag in consumption in Chapter 2; this is why we have chosen form (3) for the consumption function. Investment's lead is a result of the accelerator, that is, the level of investment is linked more to the change in production than to its level. In fact, if the acceleration effect were the only thing to consider, investment would be proportional to the speed of change of production; the result would be a sine curve shifted to the left by one quarter of a cycle with respect to the path of production. Equation (18) as chosen, features long lags which explain why the length of investment advance is much less marked.

It is easy to understand the dynamics of the cycle. During the depression, investment is low; consumption, which at first sustains demand, later falls progressively. But before production reaches its minimum level, investment recovers somewhat, for the change in production is no longer so low and the change in capital requirements are not so depressed. Recovery in investment brings with it a recovery in production which speeds up for a while before it slows down again.

The description above illustrates the theory explaining economic cycles by the interplay of the accelerator and the multiplier. It does not claim to apply exactly. We shall see later that the lags actually observed are less manifest than they appear to be on Figure 1. In particular, the lead of investment with regard to production plays a crucial role here but does not appear systematically in empirical evidence. We already noted the fact and explained it in Section 5.1 of Chapter 4 when considering the acceleration phenomenon for fixed investment. We shall come back on these empirical issues, first in Section 2.2 and later in the second part of this chapter.

1.3. Limitation of productive capacity

Behaviour which affects the formation of aggregate demand thus seems to result in cyclical fluctuations. The above analysis, however, assumes that supply constantly adapts to demand, without imposing any constraint on the latter. This is obviously a simplification, and we must now come back to it. Indeed, we expect that, in phases where expansion is especially rapid, the degree of utilization of productive capacity rises up to such a level that any additional increase in output then becomes impossible. At the very least, if an additional increase occurred, it would give rise to inflationary pressures, which would interfere with the behaviour specified so far.

In other words, a relatively regular and balanced economic growth will be accompanied by a rate of capacity utilization which oscillates around a "normal" level, inside a fairly narrow band in which no particular pressure appears. The model in the preceding section so applies as long as the time path obtained remains within the band in question. The model has to be revised when we want to deal with periods in which the degree of capacity utilization becomes abnormally high. Such periods can appear, either as a result of the dynamics of induced demand, or even more simply, as a result of the numerous movements in autonomous demand A_t.

To understand the phenomenon, we are going to consider it first by means of a deliberately simplified model[12]. Then we shall briefly introduce one by one some important complications which can affect it.

Let us rule out irregularities in autonomous demand A_t and rather assume that it increases at the constant rate g of long-run growth, a sort of "natural" growth rate:

$$A_t = A_0(1+g)^t. \tag{25}$$

[12] The model is essentially that chosen by J. Hicks, *A Contribution of the Theory of the Trade Cycle* (Clarendon Press, Oxford, 1950).

If the initial conditions would allow, production could thus grow steadily at the same rate:

$$y_t = \hat{y}_0(1+g)^t. \tag{26}$$

Indeed, such an expression constitutes a particular solution of Equation (19) when $\hat{y}_0$ is chosen appropriately, that is, when it is defined by:

$$\begin{aligned}&\bigl[1-\alpha(1-\gamma)\bigr](1+g)-\bigl[\gamma+b(1+\delta-\gamma)\bigr]+\frac{b(1-\gamma+\gamma\delta)}{1+g}\\&\quad=\frac{A_0}{\hat{y}_0}(1+g-\gamma)\end{aligned}$$

an equation which generalizes (24) and thus specifies the long-term multiplier when growth occurs at a non zero rate.

Let us also assume, in addition, that productive capacity increases steadily at the same rate g and that balanced exponential growth (26) would correspond to a normal rate of capacity utilization; the margin M of unemployed capacity would thus be constant. This is the same as saying that productive capacity $\bar{y}_t$ grows according to[13]:

$$\bar{y}_t = \hat{y}_0(1+g)^t + M. \tag{27}$$

As we have already done elsewhere, we are going to assume that output y_t depends only on aggregate demand (and that this is determined by the simple equations set out above) so long as y_t remains below $\bar{y}_t$. On the other hand, if aggregate demand tends to increase to a level above $\bar{y}_t$ there will have to be rationing, since supply will then hit its upper limit $\bar{y}_t$. This limit may be said to fix a "celling" on the growth of output.

The model of the multiplier and accelerator, which we examined in the previous section, will thus apply so long as its solution will satisfy:

$$y_t \leqslant \bar{y}_t.$$

A damped sine curve, or even one with a constant amplitude, can so be prolonged indefinitely when the initial conditions give it an amplitude that does not exceed M and when autonomous demand evolves according to (25). But

[13] The margin M appears here in an additive form rather than in a multiplicative one in order to simplify what follows. The inelegance of the equation has little practical importance since here we are studying short-run movements.

the situation will differ if either the reference time path has amplified cycles or, what is more likely, if irregularities in autonomous demand frequently modify the actual path and sometimes gear the evolution of aggregate demand toward fluctuations of amplitude greater than M. The actual path will thus hit the ceiling when aggregate demand will run up against the supply limit.

It should be understood that this has a rather profound effect on the subsequent evolution. We cannot simply say that the upper part of a path of the sort shown in Figure 1 is replaced by evolution along the ceiling: actually the whole path will take a different course from the one it would have followed, had it not been limited by the ceiling. Each shock against the ceiling sends the path off in a new direction, making it start, as it were, from new initial conditions.

In economic terms the rationing, imposed by the productive capacity limit, means a break which affects not only production, but also the incomes created by this production and hence, the demand which comes from these incomes. It is easy to understand that, since the increase in income is slowed down just at the time of a boom, the shock against the capacity limit will throw economic evolution on to a fluctuation which is likely to have a lower amplitude than that which prevailed previously. We may, in fact, check this by using the mathematical model we examined in the previous section.

If autonomous demand evolves according to Equation (25), then output satisfies the equation:

$$y_t = \hat{y}_0(1+g)^t + a\rho^t \cos(\omega t + \varphi). \tag{28}$$

So long as it does not run up against the limit of capacity $\bar{y}_t$, where it is understood that the values of ρ and ω are given by the solution of Equation (20) and that a and φ are two constants depending upon the initial conditions. We want to check that, after hitting the limit $\bar{y}_t$, output will evolve according to the new equation:

$$y_t = \hat{y}_0(1+g)^t + a'\rho^t \cos(\omega t + \varphi'), \tag{29}$$

where the positive number a' will be different from a and probably lower.

Let us consider the second terms of the right-hand side of (28) and (29). Dealing with the problem in discrete time would be unnecessarily tedious; so let us proceed as if y_t was defined and subject to the constraint $y_t \leqslant \bar{y}_t$ for all real values of t (see Figure 2). At a particular time t_0 the constraint becomes binding, that is:

$$a\rho^{t_0} \cos(\omega t_0 + \varphi) = M. \tag{30}$$

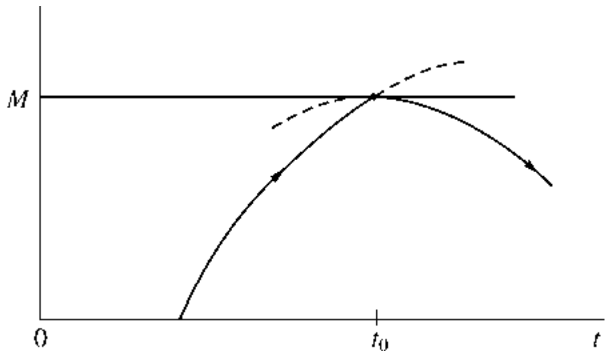

Fig. 2.

In addition the constraint tends to be violated, for without this there would be no real shock and no rationing of demand; this means that the derivative of the right-hand side of (28) is positive at t_0, hence:

$$-\omega\rho^{t_0}\sin(\omega t_0+\varphi)+\frac{\mathrm{d}\rho^t}{\mathrm{d}t}\cos(\omega t_0+\varphi)>0. \tag{31}$$

The shock of hitting the ceiling has the result that, for $t>t_0$, the time path follows a new curve which at t_0 is tangent to the ceiling. At t_0 the second term of (29) is therefore such that:

$$a'\rho^{t_0}\cos\bigl(\omega t_0+\varphi'\bigr)=M, \tag{32}$$

$$-\omega\rho^{t_0}\sin\bigl(\omega t_0+\varphi'\bigr)+\frac{\mathrm{d}\rho^t}{\mathrm{d}t}\cos\bigl(\omega t_0+\varphi'\bigr)=0. \tag{33}$$

It follows immediately from Equations (30) to (33) that:

$$a\cos(\omega t_0+\varphi)=a'\cos\bigl(\omega t_0+\varphi'\bigr), \tag{34}$$

$$-a\sin(\omega t_0+\varphi)>-a'\sin\bigl(\omega t_0+\varphi'\bigr). \tag{35}$$

These two relationships clearly imply that a' differs from a.

If $\rho\leqslant 1$, it follows from (33) that $\sin(\omega t_0+\varphi')$ is negative. So, squaring the two terms of (34) and (35) and adding them, we find indeed that $a^2>a'^2$. It is remarkable, but without any great practical importance, that this result cannot be established in general when $\rho > 1$; it is valid only if the derivative of $a\rho^t \cos(\omega t + \varphi)$ is not only positive at t_0, but also large enough.

The sketch of the argument above should clearly show that periods when capacity is fully used play an essential part in the cyclical evolution. But it is obviously too simplified to well describe what actually happens.

Note first that, according to assumption (27), productive capacity grows exogenously and steadily, without being influenced by endogenous changes in investment. Despite the considerable lags between investment and the time when plant or machinery is put into service, we cannot claim that this assumption is realistic. For a more satisfactory theory we should make capacity $\bar{y}_t$ endogenous and explain how it changes in relation to the size of past investments[14]. But this would make the model much more complex and the resulting time paths more involved. It is doubtful whether we would then understand business fluctuations so much better.

Note also that our model is too rigid in assuming that productive capacity imposes an absolute limit on output but that no difficulty appears so long as capacity is not totally used. The pressure on capacity would then play no role before the shock of hitting the ceiling. In fact, as we have seen, especially in Chapter 7, Section 2.1, reality is often mid-way between the two visions, which correspond respectively to the under-employment equilibrium where aggregate demand is meant to determine output, and the full-employment equilibrium where capacity is meant to impose the level of output and demand has to adapt, in particular through a change in prices. During the cycle the increase in prices, and the slowing down effect that it has on demand, will begin to appear before complete use of capacity is achieved. The time path will be less bumpy than the one shown in Figure 2 and the limitation of demand may prevent actual rationing.

Moreover, factors other than the physical limitations of capacity can affect business cycles, and they were ignored in the previous section and in this one. Basically, we were content with making the theory of the multiplier dynamic. Now, we saw in Chapter 7 that, in order to lay the foundations for multiplier theory we have to assume rigid prices and interest rates. If this rigidity does not hold throughout the cycle, things are more complex than we have assumed. We shall come back to this.

To conclude these remarks, let us still state that "floor" effects may possibly play a similar role to the "ceiling" effects considered here. Above all, they might affect the time path of investment. Equation (4) which we have

[14] To see how the model can be set up and dealt with in this way, the reader may refer to A. Smithies, Economic fluctuations and growth, *Econometrica* (January 1957).

used here, may in fact lead to unrealistically low, or even negative, values of investment in a time of depression. In some cases of real slow-down the fall in investment, and hence in production, may so be less pronounced than the model implies; this also would affect the subsequent evolution.

Mathematically, the fact that there is a ceiling, be it rigid or not, means that the path of the economy cannot be well represented by the solution of a complete system which would be purely linear. The theory of linear dynamic systems, the elements of which are to be found in Section 1.2, does not apply to non-linear systems. Even though some early works in mathematical economics had considered non-linear systems with ceilings, we shall not present them here[15]. We remember that complications resulting from non-linearity were already approached in Chapters 6 and 7. Others will appear at various places in the next part of this chapter.

1.4. The reference cycle

We are now going to come back to other simplifications which allowed us to set out the model in Section 1.2, the so called "model of the multiplier and accelerator". We then aimed at showing why fluctuations in inventories building and fixed investment induced an oscillatory time path of production. But other factors may be instrumental in creating such a path, or in altering its course. In truth, the latter are so numerous, the lags which they generate so varied, that a theory integrating all of them would get lost in complications. We shall better realize this as we proceed through Sections 2.1 to 2.10 and subsequently to econometric models of national economies.

In this part of the chapter, we are still going to look, one by one, at some of the factors which attracted our attention in the preceding chapters. The purpose will be to know how each of these taken alone would influence the time path of an economy already assumed to evolve in cycles such as those described in Section 1.2. We particularly want to see whether the factor in question acts as a stabilizer or destabilizer, that is, whether its presence would damp or magnify cyclical fluctuations. The factor will thus be treated as introducing a small perturbation which will change the path that production would have followed without it. Proceeding in this way means that we shall be able to add to what we did in other chapters, without having to construct in each case, a new "business cycle theory" and study the time paths it would generate.

[15] A model very close to those discussed in this section and the previous one can be found in R. Goodwin, The non-linear accelerator and the persistence of business cycles, *Econometrica* (January 1951).

In the next part of the chapter we shall, on the contrary, consider more fundamental revisions or even replacements of the model of the multiplier and accelerator. When later approaching econometric work on national economies we shall so be equipped with a broader spectrum of theoretical references.

At present, the basic framework will be the model in Section 1.2 in the relatively specific form which led to Equation (19). We want to see how each of the additional phenomena to be considered affects the formation of aggregate demand and so requires modifications in the rather sketchy equations previously used for consumption and investment. This analysis will so result in the introduction of additional terms in Equation (19). We shall then study whether presence of these terms increases or decreases the amplifying coefficient ρ, which was given by (22) for the paths ruled by Equation (19). To facilitate the argument, we shall often take advantage of the assumption that perturbations and additive terms are small.

More precisely still, we shall indulge in assuming a six-year basic cycle and in letting the coefficients of Equation (19) have the values suggested in Section 1.2. This will be our "reference cycle". In addition, when the occasion arises, we shall not let the additive terms introduce lags greater than 2 in Equation (19), so that the characteristic equation remains one of second degree. Economic reality is obviously less simple, but to try to show it in all its complexity is not in the spirit of our approach at this point.

Before tackling the substance of our enquiry, it will be useful to see, in abstract terms, what role the additive terms play. Suppose, then, we add perturbation H_t to aggregate demand described by the consumption function (3) and the investment function (18). Suppose also that H_t is expressed as the following function:

$$H_t = h_0 y_t + h_1 y_{t-1} + k, \tag{36}$$

where h_0, h_1 and k are constants.

It is often possible to obtain an expression of this type as a first approximation, especially after having studied the lead or lag of H_t with respect to production. So, if H_t tends to increase just when production does, we can take $h_1 = 0$ and $h_0 > 0$; if H_t has a lag of one year with respect to y_t, we take $h_0 = 0$ and $h_1 > 0$; if H_t has a lag of two years on

$$y_t = q\left(\cos\frac{\pi t}{3} - 1\right) + y_0$$

we can take $h_1 = -h_0 > 0$; if H_t has a lead of a year, we take, likewise, $h_1 = -h_0 < 0$, etc.

Introducing term (36) into the equations in Section 1.2 raises no problem in principle. We can say that the constant k is stabilizing or destabilizing according to whether it is positive or negative. Indeed, adding a positive constant k to autonomous demand A_t, which is also positive, reduces the relative amplitude of the fluctuations in demand and consequently also the amplitude of the fluctuations induced in production.

The characteristic equation (20) is modified by addition of $-h_0$ to the coefficient of u^2, by that of $h_1 - \gamma h_0$ to the bracket multiplying u and by that of γh_1 to the constant term. Expression (22) defining the amplifying factor ρ^2 has to be modified by adding γh_1 to the numerator and by subtracting h_0 from the denominator. The perturbation is thus destabilizing if h_0 and h_1 are both positive, it is stabilizing if they are both negative. If $h_1 = -h_0$ the coefficient ρ is a decreasing function of h_1 when $1 - \alpha + \alpha\gamma > b\gamma(1 - \gamma + \gamma\delta)$, which is in fact the case; hence, the perturbation is destabilizing if $h_1 < 0$ (H_t has a lead of a year), stabilizing in the opposite case (H_t has a lag of two years). It is not surprising to find that, in short, H_t is destabilizing if it is more or less synchronized with production, the lag or lead being shorter than a year and a half.

While we shall consider the main cases of application of the approach just sketched, we should remember that reference to observed facts will be somewhat loose at this stage. The text will mean to refer to what seemed to be commonly perceived as stylized facts in the 1960s. Subsequent work showed that these facts were perhaps not as robust as was then believed. The reader will find in the next part of the chapter a fuller discussion about the difficulty of establishing a firm core of such facts and about what the content of this core could now be. Already in the following pages we shall understand why conditions for the validity of some statements may involve details in the specification or estimation of the relevant behaviour. The main propositions which were held to be true in the 1960s are nevertheless still today interesting because of their simplicity and of the good first approximation they offer.

1.5. Income distribution

First of all we are now going to come back to simplifications which concerned the distribution of incomes, and allowed us directly to replace real disposable income by the volume of production. Assuming moreover consumption to depend only on the evolution of aggregate real disposable income, we could write Equation (4). We are now going to examine the two complications arising respectively from the discrepancy between output and real disposable income and from possible differences between the marginal propensities to spend various incomes.

1.5.1. *Automatic stabilizers*

Earlier, we considered the link between disposable income and production. We posed, first in Chapter 1 and then in Chapter 7, the accounting framework which, although admittedly simplified, was acceptable for theoretical analysis. Let us look at it again with the following equation for disposable income Y_t:

$$\frac{Y_t}{p_t} = \frac{S_t}{p_t} + (1 - \tau_t) y_t, \tag{37}$$

where τ_t is the tax rate in period t, and S_t is the value of transfer payments made by government to households. These two values are considered as resulting from public decisions, and thus as exogenous.

We can easily see how taking the two variables into account changes the results reached in Section 1.2. The behavioural equation (3) and definition (6) meant that we had to express the volume of consumption by:

$$C_t = C(L)\frac{Y_t}{p_t} + \frac{c}{1-\gamma}.$$

We then derived simply:

$$C_t = C(L) y_t + \frac{c}{1-\gamma}. \tag{38}$$

Taking now more exactly the dependence of disposable income on production into account with Equation (37), we arrive at:

$$C_t = (1 - \tau_t) C(L) y_t + C(L)\frac{S_t}{p_t} + \frac{c}{1-\gamma}. \tag{39}$$

The implication is clear for the central equation (5), which expressed how output was dynamically determined. We now should add the positive term $C(L)S_t/p_t$ to autonomous demand A_t and we should replace $C(L)$ by $(1 - \tau_t)C(L)$.

It is easy to see that these changes would dampen the intrinsic fluctuations if fiscal policy was confined, in a purely passive way, to maintaining constant values of the instruments τ_t and S_t/p_t. Indeed, as we have seen, adding a constant positive term to autonomous demand, which is already positive, reduces the relative amplitude of the fluctuations induced in production. Moreover, multiplying $C(L)$ by the constant $1 - \tau$, which is positive and less than 1, is the same as replacing α in Equation (6) by the smaller number $(1 - \tau)\alpha$, which should result in a damping of the intrinsic time path that serves as a reference here.

To check this last point, let us consider the case where the investment equation is given by (18)[16]. In the characteristic Equation (20), we see straight away that the sum and the product of the roots are increasing function of α and we can check that, if both roots are real, the largest is also an increasing function of α. Replacing α by $(1 - \tau)\alpha$ results in a reduction of the modulus of the two complex roots or in a reduction of the largest real root. Hence, the original time path of y_t is damped. In the case where this path is cyclical, the period of the cycle is reduced, since the ratio of the sum of the roots to their product is equal to $2\cos\omega/\rho$, in the notation of Section 1.2, and is independent of α.

To sum up, replacing (38) by (39) stabilizes the path of output y. This is called an "*automatic stabilizer*" effect, a result of the way in which public finances operate, through tax laws and decisions about transfer payments. In a period of recession, Equation (38) would imply that the decline in real disposable income be equal to that in output; Equation (39) takes account of the fact that part of the fall in incomes is borne by the public budget; this part is even more than proportional because there are transfers from which households benefit and whose amounts are independent of business conditions[17].

The other side of the coin is that the surplus on the government's current operations fluctuates with the business cycle. The real value of this surplus is equal to

$$\tau_t y_t - \frac{S_t}{p_t} - G_t.$$

If fiscal policy is purely passive, τ_t, S_t/p_t and G_t can be considered as constant, so that the real surplus changes with production. If government's current operations are in equilibrium on average during the cycle, deficits will appear during depressions, and surpluses during booms.

A fiscal policy which would regulate taxes and transfers so as to continually ensure budgetary equilibrium would obviously run counter to the automatic

[16] This is not exactly the case we dealt with in the previous section since H_t now has the form $-\tau C(L)y_t + k$.

[17] In reality, we should take into account the lags in collecting taxes, the fact that some of these are progressive and the fact that some transfers increase during cyclical downturns (unemployment benefits in particular). On the whole, this reinforces the importance of automatic stabilizers.

stabilizers. This is the reason why this policy is not often recommended. In some countries, governments are even asserting that they are aiming at equilibrium for the "full-employment budget". They are then publishing an amount such as:

$$\tau_t p_t \bar{y}_t - S_t - p_t G_t,$$

which would have been obtained for the budget surplus in year t if output had reached the full-employment level $\bar{y}_t$.

1.5.2. *Fluctuations in the labour share*

To assume that the path of consumption depends only on that of aggregate disposable income rules out the effects of possible changes in income distribution. Likewise, to assume that the path of investment depends only on the need for productive capacity rules out possible changes in firms' financial constraints, changes which are obviously linked to the distribution of incomes.

In order to examine how the results above are affected by these thus far neglected complications, we shall characterize the distribution of incomes by just the "labour share", which may be estimated from national accounts as the ratio to GDP of wages and other labour costs. In our notation we shall write:

$$s_t = \frac{w_t L_t}{p_t y_t}. \tag{40}$$

In our discussion we shall, now like in Section 1.2, forget about the possible interference of public finances, which we have just dealt with. In order to proceed we shall first consider how the labour share fluctuates, then, how these fluctuations affect the behaviour equations we specified for consumption and investment.

Actual fluctuations in the labour share are empirically known to be countercyclical: the labour share increases particularly fast in the first phase of the recession and usually peaks more than a year after the deviation of output from its trend. It is one of those business cycles stylized facts which we shall consider more systematically in the next part of the chapter. We even know that systematic movements of the labour share may often be decomposed into two systematic movements of respectively the real wage rates and labour productivity. Considering deviations from trends, we know of many cases in which the real wage rate, w_t/p_t, or more precisely the real labour cost *per* unit of output[18], peaked with a lag behind production. On the other hand fluctuations in the average productivity of labour, y_t/L_t, quite systematically "lead" (really precede) those in production.

The procyclicality of the real wage rate, but with a lag with respect to output, is consistent with simple expressions of the theoretical and empirical material discussed in Part 6 of Chapter 8. But the connection is fairly involved.

In Section 6.2 of that chapter the rates of change g_{wt} of the nominal wage rate and g_{pt} of the price level were related to indicators of demand pressure, respectively in the labour market (u_t) and in the goods market (d_t). Deriving equations for the rate of change $g_{wt} - g_{pt}$ of the real wage is easy. With the general formalization posed in Section 6.2 it led to Equation (296) for the case in which other exogenous variables than the pressures of demand could be neglected, a case that suffice for our present discussion. With the particular specification (243) the result is:

$$\begin{aligned} g_{wt} - g_{pt} = {} & D(L)(1 - bL)\big[(1 - c + cb - bL)\sigma(u_t) \\ & - b(1 - L)\pi(d_t)\big], \end{aligned} \tag{41}$$

where $D(L)$ was given by (245). But deriving such an equation is just a first-step for establishing the dynamic pattern linking the real wage rate to output.

We defined in Chapter 8 the pressure of demand on the goods market in such a way that it is synchronous with output (see Equation (227) there and note that productive capacity appears in the equation, whereas D_t is meant to be equal to output y_t). But the unemployment rate u_t, the indicator of demand pressure on the labour market, lags behind output (this is about the same phenomenon as the lead of labour productivity with respect to output, which we shall consider in a moment). For simplicity we took advantage of the fact that strong pressure on the goods market usually corresponds to strong pressure on the labour market and posed Equation (229) which determines u_t as a lagged function of

[18] Identifying the real wage rate with the real labour cost per unit of output is not always suitable, particularly in applied macroeconomics. The reader may find the distinction being made between "the product wage" and "the consumption wage", both concerning specific definitions of the ratio w_t/p_t, seen from the viewpoint of either the employing firm or the consuming household to which the employee belongs. In the product wage, w_t is meant to be the labour cost of a unit of labour for employers and p_t the price of a unit of output (actually a unit of the volume of value added when aggregates are considered). In the consumption wage, w_t is meant to be the net wage rate received by the worker and p_t the price of a unit of consumption. The product wage and the consumption wage evolve in proportion to one another under two conditions: fixed rates of taxes and of related mandatory contributions, unchanged terms of foreign trade. These conditions will be assumed to hold in what follows, but are seldom exactly met in applications, which will then require due adaptations.

the sequence of d_t. Accepting again this equation we are thus led to a dynamic relation directly linking the sequence of $g_{wt} - g_{pt}$ and to the sequence of d_t.

But non-linearity of the two functions $\sigma(u_t)$ and $\pi(d_t)$ still leaves us with a potentially complex relation. Indeed, in Section 7.2 of Chapter 8, when considering reactions of the real wage rate to changes in demand, we had to distinguish various cases depending on the relative initial levels of pressure of demand in the two markets. Assuming, as we do now, that an increased pressure induces an increase in $g_{wt} - g_{pt}$ is therefore ruling out cases that may, and at times do, occur. (This remark among others calls to mind that empirical regularities qualified as stylized facts are not always robust. Precisely on this case of real wage rates we shall in Section 2.2, at the end of Subsection 4, refer to the ambiguous evidence of procyclicality.)

Establishing why average labour productivity should lead output is easier because it is a direct consequence of the presence of adjustment lags discussed in Section 1.3 of Chapter 4. We there saw that actual employment is likely to be a direct function of the demand for goods (see conclusion at the end of Section 1.2 of that chapter). We also examined in Section 1.4 the econometric evidence about the short-run fluctuations in the demand for labour; we then related employment, the length of time worked and labour productivity to output, also concluding that productivity should be slightly leading output. Such a conclusion was found to be consistent with the ones drawn by D. Hamermesh from his extensive survey of econometric investigations about the demand for labour.

Given this multifarious evidence, we can well understand the empirical evidence showing that the share of wages in total income has, with regard to production, a substantial lag. This lag well exceeds a year, to the point that the labour share often looks simply counter-cyclical.

1.5.3. *Stabilizing effect via consumption*

It is commonly argued that a given aggregate disposable income leads, all other things being equal, to consumption which is all the higher as the labour share in that income is higher. So, in the consumption function, we should distinguish income from labour from other incomes, or at least we should include the labour share as a corrective factor alongside total disposable income. Let us look at the consequences, reminding however that the corrective factor is not very large (see Section 2.8 in Chapter 2).

Assume that the consumption function (3) is replaced by:

$$C_t = \alpha_1(1-\gamma)\frac{w_t L_t}{p_t} + \alpha_2(1-\gamma)\frac{P_t}{p_t} + \gamma C_{t-1} + c, \tag{42}$$

where P_t denotes income other than what is accruing to labour[19]. This is approximately the same as adding the term:

$$H_t = \frac{(\alpha_1 - \alpha_2)}{1 - \gamma L} \overline{Y}(s_t - \bar{s}) \tag{43}$$

to aggregate demand, where it is understood that α has then to be equal to $\alpha_1 \bar{s} + \alpha_2(1 - \bar{s})$, whereas $\bar{s}$ and $\overline{Y}$ denote the values, on average during the cycle, of s_t and Y_t/p_t.

Expression (43) shows that H_t has a lag of about six months with respect to the labour share s_t, which has, with respect to output, a lag definitely exceeding one year. So H_t is almost "countercyclical". Hence, through consumption, fluctuations in the labour share play a stabilizing role. In phases of depression, the slowing down in demand is lessened by the fact that the distribution of incomes goes relatively more to those categories of households which have a high propensity to consume.

1.5.4. *Destabilizing effect via investment*

But the effect on investment works in the opposite direction. In Part 4 Chapter 4 we saw that availability of internal funds and ease with which credit is obtainable might positively affect the level of investment. This was confirmed to be the case in Section 5.2 of Chapter 4. Now, financial easiness is at its maximum in the middle of a boom, that is about one year before the maximum of the output cycle. The expansion of investment is thus reinforced, which increases aggregate demand exactly when it tends to be at its highest.

We can see that financial easiness is particularly high in the middle of a boom by looking at different statistical indicators. For example, the proportion of executives who say they experience financial difficulties is lowest before expansion reaches its highest point, and is highest before the depression reaches its lowest point. We come to the same conclusion when considering fluctuations in the labour share and in the rate of self-financing.

As was shown in Chapter 4, Part 4, financial easiness is linked to the structure of firms' balance sheets and to credit rationing. Firms' balance sheets tend

[19] Function (42) can be criticized because it assumes the same lags for consumption spending whether this be from wages or other incomes. An alternative assumption would be to suppose that all disposable incomes of households are spent, but at different speeds. So we would go back to assumptions made in Chapter 8, Section 7.1. By analogy with Equations (294) and (295) in Chapter 8, one could, for example, obtain a corrective term H_t expressed as:

$$(b_P - b_W)(1 - L)(s_t - \bar{s})/(1 - b_W L)(1 - b_P L) \quad \text{with } b_P > b_W.$$

The difference with (43) is of little importance if b_P is close to 1.

to be particularly favourable at the end of a period during which the rate of self-financing, measured as a ratio to investment, was high, that is, at the end of a period of recovery, as we are going to see. Moreover, credit rationing only affects things at the end of a boom when the demand for credit, having accelerated progressively, reaches a very high level.

Statistical observations show that the rate of self-financing has a certain lead with respect to production and even to investment. It diminishes before the turn-around in production, when growth of investment is faster than that of retained earnings; the rate of self-financing similarly turns up before the recovery.

A time path of this sort is compatible with those we attributed earlier to investment and to the labour share. Indeed, let us denote retained gross profits by P_{rt} and hence the rate of self-financing by $P_{rt}/p_t I_t$. Note that distributed profits amount to a more or less constant share of production, whereas on average wage income is about three times higher than retained gross profits. Hence, the growth of the rate of self-financing is determined by:

$$g(P_{rt}/p_t) - g(I_t) = g(y_t) - 3g\left(\frac{w_t L_t}{p_t y_t}\right) - g(I_t), \tag{44}$$

where $g(x)$ denotes the growth rate of the variable x. This suggests that the rate of self-financing is low when the labour share is high.

An explicit calculation can make this point precise. Keeping the assumption that investment has a lead of six months over output and the labour share a lag of two years, taking into account that the growth rate of investment fluctuates with a relative amplitude about double that of output, which is about one and a half times that of the labour share, we can determine, with calculations which we shall not reproduce here, that the rate of self-financing evolves with a lead of about five quarters with respect to output.

Thus, firms' financial easiness usually ceases to improve before investment reaches its maximum. This factor reinforces the role of the accelerator: towards the end of the boom, firms slow down their decisions to invest, not only because their need for productive capacity becomes less urgent but also because financing becomes less easy to obtain. Just as the accelerator explains fluctuations in output, so the reinforcement of the accelerator plays a destabilizing role.

1.6. Prices and interest

Factors other than capacity needs, financial easiness and the differences between marginal propensities to consume affect the development of aggregate demand. In particular the wealth effects and changes in the relative costs of capital and labour ought to play a role. In order to deal with them we must, first of all, study fluctuations in the general price level around its trend.

1.6.1. *Price level and wealth effects*

In Chapter 8 we examined in great detail the links between production and the growth rate g_{pt} of the price level, established between t and $t+1$. Putting our conclusions into the six-year reference cycle, we can see that fluctuations in the rate g_{pt} (observed in $t+1$) have a slight lag with respect to those in production, which conforms to direct observation of the statistical series, at least if we eliminate the effect of exogenous shocks whether deliberate or not (in France for instance, large wage increase in June 1968, increase in the price of oil in January 1974 and in 1979–80, its decrease in 1985–86, the wage policy decided in 1983, ...)[20].

In Part 2 of Chapter 8 production was characterized by the variable q_t (see Equations (32) in Chapter 8). The equations which were subsequently set out established a direct link between fluctuation in q_t and those in g_{pt} (see, for instance, Equation (48) in Chapter 8). With the numerical values from the estimated Equation (57), based on US data for 1953–1978, and with a six-year cycle, the average lag in the rate of inflation with respect to q_t would be of five months (the calculation is not reproduced here and we should remember that g_t denotes the rate of inflation between t and $t+1$). The lag in the rate of inflation with respect to production is also found in less aggregated models. We considered this especially in Section 7.1 of Chapter 8 when we included the rate of increase in wages alongside the rate of increase in prices.

Changes in the rate of inflation act first through wealth effects. The purchasing power of wealth formerly accumulated by households in the form of cash balances or bonds is lowest when the difference between the price level and its trend value is highest, which happens a quarter of a cycle after the maximum of g_{pt} (the level of a variable has a lag of a quarter of a cycle with respect to the rate of change of that variable). Wealth effects will thus depress households' consumption during recession and will stimulate it during expansion. Even though the lag of p_t behind y_t is not much longer than a quarter of a cycle, the wealth effects on consumption will be destabilizing. As the price increase strengthens firms' financial positions, effects on investment, on the contrary, will be stabilizing; but we concluded earlier that wealth effects are more important for households than for firms.

[20] Like all other stylized facts quoted here, this suffers from numerous exceptions. For an analysis of English and American cycles, see E. Shapiro, Cyclical fluctuation in prices and output in the United Kingdom, 1921–71, *Economic Journal* (December 1976).

1.6.2. *Inventories and the real interest rate*

While the rate of increase in prices g_{pt} has a slight lag with respect to production, the rate of expected inflation g_{pt}^{e} usually has a more substantial lag. The expressions in Chapter 8 suggests a lag of about a year, which would apply in the absence of a sharp change in expectations or exceptional exogenous influences.

These fluctuations in the expected rate of price increase can affect agents' stocking decisions: households may buy durable goods sooner or later, firms may decide an increase or decrease in the speculative component of their stocks (that is, in the amount of inventories, after correcting for involuntary accumulation or depletion beyond what the accelerator alone would imply).

In fact, depending on the case, the relevant variable will be either the expected rate of price increase or the real rate of interest. Everything depends on the terms of arbitrage: the expected rate of price increase is important for those who have to chose between stocking and holding on to their cash (which is the case for quite a few households). The real interest rate is important for those who finance their stocks by outside credit or who have to chose between inventories and financial investment on the stock exchange (the case of most firms).

Holding stocks by households for whom g_{pt}^{e} matters plays a destabilizing role in that the desired level of stocks is highest when g_{pt}^{e} has past its maximum, about a year after the high point of the cycle; stock accumulation is then greatest when the desired level of stocks increases fastest, that is at the time of a boom.

To know what the effect of inventories building is for those who are interested in the real interest rate, we obviously have to take into account fluctuations in the nominal interest rate (here, a short-term rate). These fluctuations are strongly influenced by monetary policy, which acts on r_t either directly, or indirectly through control of the money supply or the amount of credit. Thus, there is no question here, even less than in other cases, of stating any absolute regularity. Nevertheless, observing the past suggests that changes in the short-term nominal interest rate have usually been parallel to those in the expected rate of price increases. So, it seems difficult to conclude at this stage that any systematic effect would affect the cyclical path of demand because of the influence that changes in the real interest rate would have on stocking. (Fluctuations in inventories were already discussed in Section 5.5 of Chapter 4 and will be considered again in Section 2.7 of this chapter.)

1.6.3. *The relative costs of labour and capital*

We saw, in Chapter 4, that investment decisions should take costs into account, investment increasing as the cost of labour rises with respect to the cost of capital. It is rather difficult to specify how this consideration affects the course of the business cycle, even though a role is often given to it in discussions about the determination of short-term equilibria.

One of the difficulties stems from the lag in investment demand with respect to the decision to invest. It appears that this lag cannot be ignored, since it is no longer a matter of taking into account credit availability, which could be responsible for accelerating or slowing down the actual implementation of investment decisions. It is a matter of choosing equipments designed with the aim of economizing on labour. When we were dealing with the accelerator and justifying it in terms of the need for productive capacity, we finally chose Equation (18) in which the lag structure was represented by the operator $L/(1-\gamma L)$ supposed to apply to annual series. This operator, applied to the reference cycle with the value $\gamma = 1/4$, yields a lag for investment demand of about five quarters with respect to the factors which motivate it; but it was introduced more for convenience rather than because there would have been strong econometric justifications for it. In fact, available econometric evidence suggests that, as to the substitution of capital for labour, the lags are longer than with respect to changes in productive capacity, the need for which is more immediately perceived and concerns inventories as much as fixed investment. Hence, effects of changes in the relative cost of labour with respect to capital are likely to be spread out over a fairly long time and to play a more important role in the trend growth rate of demand than in its short-run fluctuations[21].

Let us, however, push the investigation a little further and briefly consider the fluctuations in the ratio $w_t/\rho_t p_t$ of real wages to the real interest rate (other elements of the relative cost of labour with respect to capital do not seem to be systematically linked to cyclical fluctuations). When discussing the dynamic pattern of the labour share, we saw that the growth rate of the real wage tends to have a slight lag with respect to production. This is the most significant factor here since there is no clear cyclical fluctuation in the real interest rate. The long-term real interest rate is now at stake; but the same lack of correlation with output that we noted earlier for short rates seem to apply to long ones.

Desired capital is linked to the relative cost of labour with respect to capital. Investment, therefore, must be linked to the rate of change of this relative cost.

[21] Econometric studies seem to have confirmed this idea since the early work of C. Bischoff, Hypothesis testing and the demand for capital goods, *Review of Economics and Statistics* **51**, No. 3 (1969).

If a lag of five quarters occurs between the decision to invest and its implementation and if the rate of change in question has itself a lag of one quarter with respect to the output cycle, the main effect will appear a quarter of a cycle after the high point of economic activity. It is neither stabilizing, nor destabilizing. So, this confirms the idea that changes in the relative cost of labour with respect to capital play no systematic significant role in business cycles.

1.7. The money market

Pursuing at length now the tentative exploration of leads and lags implied by aggregate demand analysis, as conceived in the 1960s, would quickly become unconvincing. It is, however, interesting to still wonder whether the requirement of the monetary equilibrium is likely to affect cyclical properties which were so far based mainly on consideration of the goods market and of income distribution. The question was approached in the 1950s, for instance by H. Minsky[22]. In the same spirit we shall here admit that the money market determines the nominal interest rate. We shall look, first, at fluctuations in the demand for money, second, at the implications following from the assumption of an exogenous and fixed money supply, third, at possible induced fluctuations of this supply.

1.7.1. *The demand for money*

In Chapter 2 we discussed models of the money demand function (see in particular Section 4.6 about empirical evidence). Here we shall limit attention to the simplest of those models which take into account the dynamic characteristics of the demand for money, namely to:

$$M_t = p_t y_{pt} L(r_t), \tag{45}$$

where y_{pt} is real "permanent income" and L is a decreasing function tracing the effect of the preference for liquidity[23] (this model chooses an elasticity of 1 with respect to permanent income, which is an acceptable approximation). Permanent income, which already appears in the consumption function, is easily

[22] H. Minsky, Monetary systems and accelerator models, *American Economic Review* (December 1957).

[23] The function L is unlikely to be confused with the lag operator, which we shall continue to also write L.

connected to output if we keep to the simplifications assumed at the beginning of Section 1.2

$$y_{pt} = \frac{\beta}{1 - \gamma L} y_t. \tag{46}$$

For the characterization of cyclical shifts in the money demand function we must then consider the cyclical behaviour of the nominal permanent income. Later on we shall have to wonder whether the nominal interest rate is maintained roughly constant along the cycle, a feature which would be congruent with the elementary multiplier theory and would make the quantity of money move in proportion to nominal permanent income. Since we realized in Chapters 7 and 8 that such an elementary theory is unlikely to be fully right, we shall have to take the requirement of the monetary equilibrium into account, our static reference then moving from the multiplier model to something like the IS-LM model.

From the previous discussion it follows that the real permanent income should have a lag of from one to two quarters with respect to output, and the price level a lag of about seven quarters. Fluctuations in the price level have often been said to be relatively more important than those in output[24]. If so, we can conclude that $p_t y_{pt}$ has a lag of one year or more with respect to y_t. So, even if there were no feedback from the money market to the market for goods and if the adaptation in the quantity of money were purely passive, this quantity would exhibit a cyclical behaviour. Hence cyclical fluctuations in the quantity of money are not sufficient evidence for showing that changes in the money supply play a causal part in business fluctuations. This is a trivial point after the long discussion we had in Section 5.4 of Chapter 8, but a point worth remembering at this stage.

1.7.2. *Stabilization via the liquidity preference function?*

In order to look at the feedback from the money market let us first assume that the money supply is somehow kept constant while the demand for money fluctuates[25]. Reference to the static models of Chapter 7 might lead us to think that induced changes in the rate of interest will stabilize the fluctuations in output and the price level. But a closer scrutiny of lags shows that the validity of such an effect may be questioned.

[24] Actually, we shall see in the next part of the chapter that this is far from being universally true.
[25] These hypothetical constancy on the one hand, fluctuations on the other hand, are to be understood as concerning deviations from trends. The remark applies more generally in this chapter, but will not be often recalled since it is rather obvious.

Indeed, the IS-LM reference in Sections 2.3 and 2.5 of Chapter 7 is illustrated by Figures 3 and 6 of that chapter, for the two cases of respectively a slack of demand and quantity adjustments on the goods market, or a pressure of demand on capacities and price adjustments. Section 1.2 of this chapter suggests that the position of the IS line on the graphs of the two figures fluctuates back to the left during depressions, forth to the right during booms. Actually the model of Section 1.2 assumes that the equilibrium varies as illustrated by a movement from P to P_1 in Figure 3 during the depression, from P_1 to P during the boom. The point of the IS-LM argument is to stress that the requirement of the monetary equilibrium should not be forgotten: because of induced changes in the rate of interest while the money supply remains fixed, the movement in output (alternatively in the price level) will have a smaller amplitude, not $[P_1 P]$ but rather $[P_2 P]$, which will be significantly smaller except in the case of the Keynesian “liquidity trap” when the interest rate remains permanently low.

But such an interpretation of the IS-LM model in a business-cycle framework does not strictly hold because it neglects the lags in adjustments. There are two main adjustments which we may be inclined to characterize as follows: (i) on the money market fluctuations in output (or prices) lead to a comovement in desired cash balances, a movement that has to be prevented by whatever change in the rate of interest is needed, given the liquidity preference function $L(r_t)$, (ii) on the goods market fluctuations in the rate of interest induce contra-movements in aggregate demand, which mitigate the movements that had induced the change in the interest rate. This characterization is, however, risky because speaking of comovements and contra-movements becomes misleading if the lags are too long: a comovement after a lag of half the cycle period really appears as a contra-movement. We must therefore have some evaluation of the lags after which adjustments (i) and (ii) above materialize.

The lag in the first adjustment is precisely that with which the demand for money adapts. We concluded a moment ago that it amounted to one year or more. As for the lag in the effect of changes in interest rates on aggregate demand, the admittedly somewhat speculative discussion of the foregoing Section 1.6 about the effects of changes in the real interest rates and about the connection between nominal and real interest rates would lead to the conclusion of a rather long lag. Everything considered, maintaining a fixed money supply would not seem to much contribute to the stabilization of cyclical movements in aggregate demand; according to this argument the effect might even conceivably be somewhat destabilizing.

Two kinds of objections may, however, be raised against the argument. In the first place, it loosely deals with the aggregation difficulty when it, implicitly, focuses on mean lags, which are not the most relevant in the present context. In order to see the point, suppose that there are two effects of changes in interest rates on aggregate demand: a short-run and a long-run effect. The short-run effect could concern voluntary changes in inventories and the timing in implementation of investment decisions of households and firms. It would be triggered by changes in the short-term interest rates, the ones which quickly react to the first changes in the perceived permanent income and in the demand for money. The long-run effect could concern the input mix chosen in productive equipments, which could be more or less capital intensive. It would depend on long-term interest rates of loans financing these equipments, rates which adapt slowly and partially to changes in the short rates. Clearly, the short-run effect is stabilizing: it involves lags that may be shorter than a year. On the other hand, the long-run effect is likely to be spread over a number of future years and to be so diluted as to have little effect, if any, on the time profile of aggregate demand at business cycle frequencies.

In the second places, the argument (like the standard IS-LM analysis more generally) accepts "the monetary view" in financial macroeconomics. If we recognize on the contrary the value of "the credit view" and the role of credit rationing, we may argue that keeping a constant money supply while the demand for money is fluctuating leads to harden credit rationing when the money demand increases and to relax rationing when this demand decreases. Since credit rationing may be assumed to act quickly the credit view, by displacing attention from the money market to the credit market, strengthens the argument in favour of a stabilizing role of these markets.

1.7.3. *Destabilization via the money supply?*

A third reference must be added to those used up to now in this section: first, a passive money supply, which would validate the neglect of financial markets in business cycle theory, as is done in the multiplier-accelerator model, second, a fixed money supply imposing a significant feedback from financial markets. Indeed, since many decades a number of economists gave a determinant role to financial instability in the explanation of observed business fluctuations: increases or decreases in the supply of money and/or credit would have been the prime mover in many of these fluctuations. The idea was widespread up to the second worldwar, as we shall more precisely see in Section 2.10. Although definitely less fashionable in the 1950s and 1960s, the idea never fully vanished.

The idea may fit into the IS-LM framework: we just have to consider important shocks shifting the LM curve, to the right in case of increases in the money supply, to the left in case of decreases. For instance, a shock increasing

M leads to a fall in the interest rate, triggering the multiplier and accelerator phenomena. The same kind of effect occurs after a shock relaxing credit rationing.

The importance to be recognized to the idea of autonomous financial shocks may be established in two ways. Either we show that such shocks may be relatively important compared with demand shocks. Or we uncover the existence of amplifiers within the financial sphere, which explain how the initial shock may trigger an important dynamic process of its own: for instance, in the case of a positive shock, a bubble of asset prices on financial markets may develop, or in the case of a negative shock, a chain of reactions from bank to bank may amplify an initial tightening in the availability of loans to households and productive firms. Existence of such amplifiers was often asserted by economists, including J.-M. Keynes (see references given by H. Minsky and M. King in Section 2.10).

We shall not enter here into a formal analysis of the intuitive views about the independent active part that financial phenomena could play in the generation and amplification of business fluctuations. At this stage the interested reader may refer to the article written in 1957 by H. Minsky (op. cit.), in which the author reaches, through somewhat heuristic arguments, conclusions such as the following two: if the quantity of money remains constant, "the business cycle will be characterized by weaker booms than would occur with a permissive monetary system" (end of his Section II); "Autonomous or cyclically induced changes in the liquidity preference relation can change the dependence of an expansion upon changes in the money supply and therefore affect the ratio of bank debt to total assets of firms... Changes in liquidity preference seem to be destabilizing" (last part of his Section III).

2 Business Cycle Facts: New Interpretations

Emergence of the Keynesian vision of business cycles in the 1940s and 1950s led to a long and fruitful interlude in the development of ideas about economic fluctuations and about methods to be used for the analysis of these fluctuations. It provided a unification, which appeared most welcome at the time, particularly to those who were looking for precise formalizations. But earlier visions did not then fully disappear. They not only helped to give perspectives on the Keynesian vision, but also fed minor streams of research, some of which would lead in the 1970s to a resurgence of the alternative vision conveyed by "new classical economics".

Research of the last four decades indisputably enriched our knowledge, both empirically and theoretically. Although it led to a less consensual teaching than was formerly the case, although many old economists painfully resented the forms sometimes taken by presentations of the outcome of this research, future histories of our discipline will recognize a definite progress. They will stress the content of this progress rather than the confusion of ideas and ideological conflicts, which at times and places did come with it. This part of the chapter should be understood as inspired by the belief that indeed the macroeconomics of business fluctuations is now moving toward a more mature assimilation of this research and of the various intellectual currents which brought it.

Not surprisingly, the present stage is reached as the outcome of to-and-fro movements from facts to theoretical models, as the result also of the emergence of new concepts and new methods for looking at facts or building models. Particularly important in this respect was the progressive familiarization of macroeconomists with the nature and role of uncertainty in the phenomena they were considering. It will so be convenient to begin here with a discussion of stochastic approaches to the study of business fluctuations (Section 2.1).

After taking stock, in Section 2.2, of the present knowledge of business cycle facts, we shall discuss the role of price flexibility. This will lead us to consider in particular the Real Business Cycles vision, which considers fluctuations as a feature of the natural reaction of economies to supply shocks to which they are exposed; this is, in a sense, the opposite of the Keynesian vision (Section 2.3). But further studies of business cycles with price flexibility will introduce a much wider range of visions (Sections 2.4 to 2.6).

The importance of changes in inventories during the cycle was observed already long ago. We shall devote Section 2.7 to theories focusing on this fact and giving it the main part in the occurrence and development of business fluctuations. The remainder of this part of the chapter will consider in turn the roles of three kinds of phenomena which may shape business fluctuations. They will concern, first, the utilization, building and scrapping of productive capacity (Section 2.8). They will, second, be related to various degrees of unemployment persistence (Section 2.9). They will, finally, come from financial behaviours and from financial institutions (Section 2.10).

2.1. Stochastic approaches to the study of business fluctuations

We realized in Chapter 8 that the concept of shock is now important in dynamic macroeconomic analysis. By definition a shock was unexpected when it occurred. It may have been the result of an isolated event that perturbed once and for all the economic system. But the past is also often interpreted as having been marked by series of similar shocks, which occurred frequently and repeatedly, their impacts being at times in one direction at times in the opposite direction. It is therefore natural for economists to refer to stochastic processes in the representation of such series of shocks and in the study of their effects.

But the scientific practices that are most appropriate in our field for the application of this simple idea were not quickly learned. We shall trace here the first attempts back to an influential article published by Ragnar Frisch in 1933. We shall then spend some time in order to well understand why introduction of stationary processes of shocks in business cycle theory was a natural generalization of early mathematical models, but a generalization that had to be based on a precise mathematical theory of stochastic dynamics. History of econometric practices will lead us to indicate the roles played by two methodologies which will be studied later in this book: those applied respectively to structural dynamic models and to vector autoregressive representations (the VARs).

Finally, our survey of stochastic approaches to business cycle theories would be incomplete if we did not also focus attention on two important complications: the separation of long-run trends from fluctuations, the presence of non-linearities in economic phenomena.

2.1.1. *Frisch on propagation and impulse problems in dynamic economics*[26]

The macroeconomics of business fluctuations was very much discussed during the nineteen twenties. Ragnar Frish himself, who was to be become one of the few founders of the Econometric Society and the first editor of the journal *Econometrica*, was reflecting on the modes of analysis which would be most appropriate for the purpose. The problem was taking a large part in his teaching in Norway and in his discussions at international meetings.

The 1933 article was influential not just because of earlier contributions by Frisch and of his presence at the forefront of international circles in which econometrics was discussed, but mainly because it was considered to have marked a turning point in the understanding of the role of stochastic disturbances. Particularly it had explained how erratic shocks could maintain oscillations which would otherwise be quickly damped. The idea was not quite new; indeed the article refers to its exposition by K. Wicksell at a meeting of the Norwegian Economic Society in 1907; the mechanism by which irregular fluctuations might be transformed into cycles had not been studied by Wicksell; but in 1927 three mathematical statisticians had independently published about fluctuations generated by such a mechanism, the Russian E. Slutsky, the English U. Yule and the American H. Hotelling. However, because of its broader economic vision, as well as of its clear distinction between two kinds of questions, those concerning the nature of random impulses and those concerning the propagation mechanism, Frisch's article well deserves recognition in the profession.

Actually, the most interesting messages of the article do not take much space, the longest section (almost half the length of the whole text) precisely studies "A macro-dynamic system giving rise to oscillations". The system is deterministic; in modern economic language we may say that it is a "calibrated" model with "time-to-build" capacity investment and with "permanent-income-based" consumption, except that "cash-in-advance" constrains aggregate demand; in mathematical terms it is a mixed linear system of differential

[26] R. Frisch, Propagation problems and impulse problems in dynamic economics, in: *Economic Essays in Honour of Gustav Cassel* (George Allen and Unwin, London, 1933); reprinted in O. Bjerkholt (ed.), *Foundations of Modern Econometrics, The Selected Essays of Ragnar Frisch* (Edward Elgar, Aldershot, UK, 1995). Written by the editor, the introduction to the book gives useful complements about Frisch's work and his relations with others.

and difference equations. Frish works out the general solution, with numerical calculations of the trend and of the three first damped sine components.

The following section, "Erratic shocks as a source of energy in maintaining oscillations", begins with: "The examples... in the preceding sections... show that, when an economic system gives rise to oscillations, these will most frequently be damped. But in reality the cycles we have occasion to observe are generally not damped. How can the maintenance of the swing be explained?... I believe that the theoretical dynamic laws do have a meaning... But... they only form *one* element of the explanation: they solve the propagation problem. But the impulse problem remains... One way [to approach this second problem]... is to study what would become of the solution of a determinate dynamic system if it were exposed to a stream of erratic shocks that constantly upsets the continuous evolution, and by so doing introduces into the system the energy necessary to maintain the swings".

Frisch then explains how a series of random shocks (actually a white noise process) can be superimposed on a linear deterministic system and how the solution corresponding to any realization of the process can be computed as a weighted moving average of the shocks. The result of a numerical simulation is exhibited, showing how introduction of the shocks maintains the oscillations and leads to somewhat changing periodicities, although "it is reasonable to speak of an *average* period and an *average* amplitude". Frisch adds: "we get a theoretical set-up which seems to furnish a rational interpretation of those movements which we have been accustomed to see in our statistical time data". Moreover, "The solution of the determinate dynamic system only furnishes a part of the explanation: it determines the *weight system* to be used in the cumulation of the erratic shocks".

We shall see in the following pages the validity of the logical arguments underlying the vision that was so exposed by Frisch. Incidentally let us note a fact that may be of interest to historians: Frisch had some difficulty to persuade economists working close to him that "there need not be any synchronism between the initiating force or forces [, the exterior impulses,] and the movement of the [oscillating] system" (second paragraph of the article).

2.1.2. *Autoregressive process*

The simplest way to represent series of random shocks is to assume existence of a process $\{\varepsilon_t\}$ of independent and identically distributed numbers or vectors ε_t with zero mathematical expectation. The process is then introduced in the model representing the phenomenon under discussion. As we saw already in Part 6 of Chapter 8 and in Part 1 of this chapter, models used in dynamic analysis are often linear. It is therefore legitimate for us to refer here again to Equation (12) which was the object of the mathematical digression in Section 1.2. We are indeed embarking on another mathematical digression for most of this section.

Limiting now attention for simplicity to unidimensional models let us then consider the equation:

$$x_t + b_1 x_{t-1} + \cdots + b_h x_{t-h} = w_t + \varepsilon_t, \tag{47}$$

which is supposed to link the series of the two observed variables x_t and w_t with that of the unobserved random shock ε_t. The difference with (12) is clearly minor (remember that we assumed that b_0 differed from zero). This new dynamic equation could be solved so as to give values of the successive x_t as function of their initial values and of the values of w_t and ε_t. The mathematical treatment which led to the solution in Section 1.2 would again apply. The difference is only that the value of x_t would be random, before it is observed, since ε_t is itself random. Hence, the series of the x_t must be viewed as a stochastic process in the same way as the series of the ε_t.

Moreover, the linearity of (47) implies that x_t can be viewed as the sum of two values x_{1t}, which would be a solution of the equation written without the random variable ε_t, and x_{2t}, which would be a solution of the equation written without the exogenous variable w_t (the sum $x_{1t} + x_{2t}$ has to also agree with initial values). It would be tedious to work here all along with the two variables. So, we shall concentrate attention on the second one, that is, on the equation:

$$x_t + b_1 x_{t-1} + \cdots + b_h x_{t-h} = \varepsilon_t. \tag{48}$$

It will be understood without any reminder that, in any application, a non-random x_{1t} should be added to the solution of this new equation. The simplification made by substituting (48) to (47) is clearly just for expository convenience.

But we shall now add a substantial restriction, which will be reconsidered later. It specifies that the set of coefficients b_τ must be such that a stationary process $\{x_t\}$ is solution of (48). By definition *a stochastic process* is stationary if the joint probability distribution of any finite set of its values for $t = t_1 + \theta, t = t_2 + \theta, \ldots, t = t_q + \theta$ is the same as the joint distribution for $t = t_1, t = t_2, \ldots, t = t_q$. This implies in particular that such quantities as the conditional expectation $E(x_t/x_{t-1})$ or the autocorrelation coefficient between x_t and $x_{t+\theta}$ are independent of t. Given that the successive values of ε_t are mutually independent and identically distributed, the conditions on the coefficients b_τ heuristically result from what was recalled in Section 1.2.

Indeed, we saw how the roots of the characteristic equation (14), with here $b_0 = 1$, determined the nature of a set of simple solutions of the equation with no right-hand member, this set being a basis for the space of all solutions of the same equation. If and only if all the roots have modulus smaller than one, all those basic solutions tend to zero. Now, it is intuitively clear that the recursive computation of x_t thanks to (48) from initial values of this variable and from observed values of ε_t will sooner or later involve anyone among the basic solutions. The reader may then understand why occurrence in the relevant solution of a basic solution which would not tend to zero would prevent stationarity, whereas if this can never happen the process $\{x_t\}$ is indeed stationary[27].

This being given, we define an *auto-regressive process*, also written "AR process", as one ruled by a linear equation of the form (48) in which the process $\{\varepsilon_t\}$ is made of independent and identically distributed variables with zero expectated value and in which the coefficients b_τ are such that all the roots of the characteristic equation (14) have modulus less than one. An autoregressive process can be multidimensional with n-vectors, x_t and ε_t, and $(n \times n)$-matrices B_τ in an equation similar to (48). The characteristic equation replacing (14) states that the determinant of the left-hand member is equal to zero, b_0 in this member being replaced by the identity matrix. There are then nh characteristics roots, all having modulus smaller than one.

2.1.3. *Periodicities*

We must now understand why the realization of an autoregressive process may exhibit a time profile qualitatively similar to that of the deviations of an economic time series from its trend. For so doing, we shall restrict attention for a moment to second-order unidimensional autoregressive processes, i.e. to those in which $h = 2$, $n = 1$. We shall easily see in this context what meaning should be given to the concept of periodicity in stochastic processes, a concept of general interest which also contributes to showing what Frisch had in mind.

According to him business cycles were due to the reaction of the economic system to erratic impulses, which could be viewed as random. Both the nature of the process ruling how the impulses occurred and the form of their transmission were important. In particular this transmission could be responsible for the kind of periodicities observed in macroeconomic time series.

In order to be more specific here, let us loosely refer to the accelerator-multiplier phenomenon discussed in Section 1.2. It concerned the mechanism

[27] Since this heuristic argument may not appear fully convincing, the reader may want to refer to rigorous proofs, for instance, in T. Anderson, *The Statistical Analysis of Time Series* (Wiley, New York, 1971), Section 5.2.1.

through which changes in autonomous demand A_t were transmitted to output. Discussing how a single change affected output we saw that the response was likely to be a damped oscillation, the amplitude of the oscillation reflecting the importance of the change, whereas its periodicity $2\pi/\omega$ depended only on the argument ω of the two complex roots of the characteristic equation (30), whose coefficients were all given by the consumption and investment functions. A succession of erratic changes in autonomous demand would then result in a succession of overlapping damped oscillations having different amplitudes but all the same periodicity[28]. Addition of all those oscillations should lead to a somewhat erratic output series, but one which would reflect the periodicity common to all oscillations.

In order to capture how the underlying periodicity appears in the realization of the stochastic process of x_t so generated we have to refer to the spectral representation of this process. Any stationary stochastic process with finite second order moments and zero expected value can indeed be seen as a sum of sine waves of various periodicities between $-\pi$ and π, the amplitude and phase of each wave being random: this is the spectral representation[29], the spectral density being the function showing how the variance of the amplitude associated with periodicity $2\pi/\omega$ varies with ω. From any time series of observed values ruled by the process a periodogram can be computed, showing the squared amplitude of the components of various periodicities in the Fourier decomposition of the series. The smoothed periodogram provides a (good) estimation of the spectral density (if the series is long enough and) if the process is autoregressive, or more generally "linear" (see below).

A purely random process such as $\{\varepsilon_t\}$ has a constant spectral density over the interval $[-\pi, \pi]$, its value being $\sigma_\varepsilon^2/2\pi$ where σ_ε^2 is the variance of ε_t, assumed to be finite here and in what follows: in other words all real values of ω equally contribute to the oscillations or irregularities exhibited by realizations of the process. A stationary autoregressive process ruled by Equation (48) has

[28] A careful reader of Section 1.2 will notice that the oscillations resulting from the model there considered would not be independent from each other if the successive changes in A_t were so. Indeed, in our argument of Section 2.1 we substituted w_t given by (9) to the z_t of Equation (8), which really corresponded to A_t. But this complication is immaterial for the heuristic argument given here. It could easily be dealt with by reference to an ARMA process for x_t rather than to a AR process. (ARMA processes will be discussed here later.)

[29] An introduction to the spectral representation of stochastic processes can be found in Section 4 of Chapter 11 of E. Malinvaud, *Statistical Methods Econometrics* (North-Holland Pub. Co., Amsterdam, 2nd ed., 1970), among many other references.

a spectral density $f_x(\omega)$ given by:

$$f_x(\omega) = G(\omega)\frac{\sigma_\varepsilon^2}{2\pi}, \tag{49}$$

where the "gain" at frequency ω is given by:

$$G(\omega) = \left| \sum_{\tau=0}^{h} b_\tau e^{i\omega\tau} \right|^{-2} \tag{50}$$

the two vertical bars being read as "modulus of" and i denoting the imaginary number equal to the square root of -1. In particular for second-order autoregressive processes $(h = 2)$, the gain may be computed from:

$$\begin{aligned}\left[G(\omega)\right]^{-1} &= [1 + b_1 \cos\omega + b_2 \cos 2\omega]^2 \\ &\quad + [b_1 \sin\omega + b_2 \sin 2\omega]^2.\end{aligned} \tag{51}$$

Since $\cos(-\omega) = \cos\omega$ and $\sin(-\omega) = -\sin\omega$ we immediately see that $G(-\omega) = G(\omega)$, so that in order to know the variations of $f_x(\omega)$ with ω we just have to look at the interval $\omega \in [0, \pi]$.

How do periodicities appear in an autoregressive process ruled by an equation such as (48)? We may expect that the answer involves the damped sine solutions of the equation with a zero right-hand member. Indeed, let us more precisely look at the second-order case:

$$x_t + b_1 x_{t-1} + b_2 x_{t-2} = \varepsilon_t \tag{52}$$

in which moreover the characteristic equation:

$$z^2 + b_1 z + b_2 = 0 \tag{53}$$

has two complex conjugate roots with modulus smaller than one. The product of these roots b_2 is equal to the squared modulus, ρ_0^2 say, and their sum $2\rho_0 \cos\omega_0$ is equal to $-b_1$. The solution of the equation with a zero right-hand member is a damped sine wave with periodicity $2\pi/\omega_0$. Let us focus on the case in which this periodicity is equal to 6, hence $\omega_0 = \pi/3$: this was the case we selected in Section 2.1 as the possible outcome of the accelerator-multiplier phenomenon for a unit period of a year. Then we can compute, for all value of

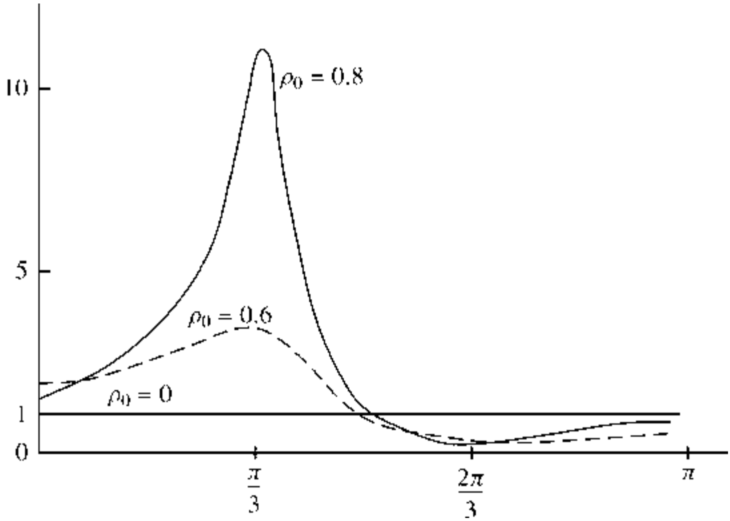

Fig. 3. Three spectral densities.

ω, the gain function $G(\omega)$ associated with (52); for so doing we can use Equation (51) with $b_1 = -\rho_0$, $b_2 = \rho_0^2$. We then see that this gain has particularly high values near ω_0. Hence, the periodic components of the spectral representation have particularly large amplitude for ω close to $\pi/3$, all the more so as ρ_0 is close to 1.

Figure 3 exhibits the outcome for the spectral density of x_t and for three values of ρ_0 respectively equal to zero (the white noise), 0.6 and 0.8 (for convenience σ_ε^2 is assumed equal to 2π).

In the interpretation remember that ρ_0^6 corresponds to the effect of dampening when we move from a wave to the next one: the ρ_0 values of 0.6 and 0.8 respectively correspond to 0.047 and 0.262 for ρ_0^6; even $\rho_0 = 0.8$ implies a serious dampening; however, importance of the periodicities around $\pi/3$ would clearly appear in a smoothed periodogram of a not so long realization of the process. The intuition according to which random shocks maintain, with some irregularities, oscillations of the series is born out.

2.1.4. *Linear processes*

Autoregressive processes do not provide a sufficiently large family for covering all stationary processes that are interesting in the study of business cycles.

Fortunately, a somewhat larger but not very different family goes a long way towards meeting the needs of this study. It is the family of *linear processes*.

A stationary process $\{x_t\}$ that has a zero expected value and is defined for all positive, negative or zero integral values of t is said to be linear if there exists a purely random process $\{\varepsilon_t\}$ and a sequence of numbers $a_1, a_2, \ldots, a_\tau, \ldots$ such that

$$\sum_{\tau=1}^{T} a_\tau^2$$

converges when T increases indefinitely and such that

$$x_t = \varepsilon_t + a_1\varepsilon_{t-1} + a_2\varepsilon_{t-2} + \cdots + a_\tau\varepsilon_{t-\tau} + \cdots \tag{54}$$

identically in t. A stationary process $\{x_t\}$ with expected value m is said to be linear if the process $\{x_t - m\}$ is linear.

We can see that an autoregressive process is linear. Indeed, if all roots of the characteristic equation (14) associated to Equation (48) have modulus smaller than one, there exists a series of numbers a_τ which meets the required conditions. For instance, with the first-order autoregressive process:

$$x_t + bx_{t-1} = \varepsilon_t \tag{55}$$

in which the absolute value of b has to be smaller than 1, the a_τ given by:

$$a_\tau = (-b)^\tau \tag{56}$$

are such that their sum converges, to $(1 + b)^{-1}$, and such that x_t and x_{t-1} defined by (54) fulfil Equation (55).

Another interesting particular case of linear processes is obtained when the series of a_τ which are non-zero is finite. We then speak of *moving average processes*, still denoted MA processes. There is a slight "abus de langage" in this common designation since, in order to define a true average the right-hand side of (54) should involve not only a finite number of terms but also only non-negative weights a_τ, which moreover should have a sum equal to 1.

The spectral density of a MA process of order k is still given by Equation (49) but with the following gain function:

$$G(\omega) = \left|1 + \sum_{\tau=1}^{k} a_\tau e^{i\omega\tau}\right|^2. \tag{57}$$

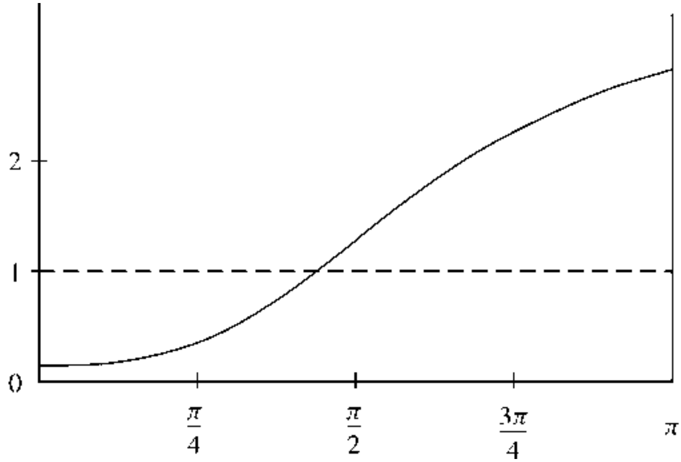

Fig. 4.

For instance, if $\{x_t\}$ is ruled by:

$$x_t = \varepsilon_t - \gamma \varepsilon_{t-1} \tag{58}$$

the gain is:

$$G(\omega) = 1 + \gamma^2 - 2\gamma \cos \omega. \tag{59}$$

Figure 4 exhibits the spectral density of $\{x_t\}$ for the case $\gamma = 0.7$ and $\sigma_\varepsilon^2 = 2\pi$. Short periodicities (corresponding to high frequencies ω) predominate; but long periodicities would predominate in case of a negative value of γ. Higher order moving averages can easily be found in which a maximum of the spectral density is reached at an intermediate value of ω like in Figure 3.

We still define a linear process, more precisely an ARMA process, if we introduce two (finite) polynomials $Q(L)$ and $R(L)$ of the lag operator L:

$$Q(L) = 1 + b_1 L + b_2 L^2 + \cdots + b_h L^h, \tag{60}$$

$$R(L) = 1 + a_1 L + a_2 L^2 + \cdots + a_k L^k \tag{61}$$

if all the roots of (14) have modulus less than 1 and if we consider the stationary process $\{x_t\}$ corresponding to the purely random process $\{\varepsilon_t\}$ according to:

$$Q(L)x_t = R(L)\varepsilon_t \tag{62}$$

(compare with Equation (8) in the mathematical digression of Section 2.1). The spectral representation of this ARMA process is still given by Equation (49) but with a gain function equal to the product of the two right-hand members of (50) and (57).

2.1.5. *Wold's theorem*

We realized, already in Chapter 8, that linear specifications are convenient for dynamic analysis. It is therefore not surprising to learn that linear processes are often assumed in stochastic economic dynamics. But we also realized, both in Chapter 8 and in Section 1.3 of this chapter, that non-linearities often have to occur in the specification of our models. So the question naturally comes to mind: how restrictive is the set of linear processes within the set of stationary stochastic process? It so happens that, in a sense, it is little restrictive. But this sense has to be precisely grasped, as we shall see later. Before stating a general result we have to introduce two new concepts.

In the first place, there are stationary processes which we might be inclined to call linear but which are not so according to the definition given above, where linear processes are related to purely random processes. Indeed, we may consider a stationary process, defined for all positive, negative or zero integral values of t, and being such that, identically in t:

$$x_t + b_1 x_{t-1} + b_2 x_{t-2} + \cdots + b_\tau x_{t-\tau} + \cdots = 0 \tag{63}$$

for a properly chosen sequence of numbers $b_1, b_2, \ldots, b_\tau, \ldots$. Such a process is called *deterministic* (a particular case, the *harmonic process*, involves only a finite sequence of non-zero numbers b_τ). Equation (63) shows that, once previous values x_θ are known (for all $\theta < t$), x_t and all the future evolution of the realization of the process can in principle be calculated; hence the term deterministic. However, randomness is still present in the amplitude or the course taken by each realization of the process (readers may like to know that deterministic processes still have a spectral representation, but no longer a spectral density). We may easily accept the view according to which stochastic but deterministic processes should not often appear in economics.

In the second place, it often happens in economics, and in other sciences, that analysts are satisfied with considering only the first and second-order moments of the random variables which they deal with. This applies also to stationary stochastic processes for which the main analytical tool often if the *correlogram*, that is, the series of the correlation coefficients r_θ between x_t and $x_{t+\theta}$ (for $\theta = 1, 2, \ldots$). Equation (54) shows that, for linear processes, those autocorrelation coefficients can be calculated as:

$$r_\theta = \sum_{\tau=0}^{\infty} a_\tau a_{\tau+\theta} \tag{64}$$

(where $a_0 = 1$). Moreover knowing the correlogram and σ_ε^2 suffices for the determination of the spectral density, which so is also a second-order characteristic. Such being the case, it is not surprising to learn that many analysts are satisfied with referring to "second-order linear" processes.

A *second-order linear process* has the same definition as a linear process except for the requirement that $\{\varepsilon_t\}$ be purely random, which is replaced by: all ε_t have zero expected value, have the same finite variance σ_ε^2 and are uncorrelated with all others $\varepsilon_{t+\theta}$. A linear process with a finite variance σ_ε^2 is second-order linear, but the converse property does not hold in general.

The following theorem, due to Herman Wold, is however worth knowing, particularly when attention is limited to second-order analysis: any stationary process $\{x_t\}$ possessing a periodogram can be expressed as the sum of two mutually uncorrelated processes $\{y_t\}$ and $\{z_t\}$ such that $\{y_t\}$ is second-order linear and $\{z_t\}$ is deterministic. At the end of this section we shall come back to the limitations to which applications of the theorem are subject.

2.1.6. *A short preliminary survey*

After the preceding long mathematical digression let us briefly survey what have been up to now the main stochastic approaches to the study of business fluctuations. We may be brief at this stage because developments given subsequently will hopefully familiarize the reader with these approaches.

We saw since the beginning of this chapter how, in the middle of the twentieth century, the theory of business cycles appeared to directly proceed from aggregate demand analysis, more precisely from the accelerator-multiplier model. Time series of consumption, investment and output were explained by time series of autonomous demand. To this deterministic approach was easily added, as a natural complement that had been recommended by Ragnar Frisch, the idea of a stochastic process of shocks to aggregate demand, shocks coming from unexplained changes in the behaviour of consumers and investors, or in foreign demand. The theory of linear stochastic processes was well fitting within the basic model.

Essentially the same stochastic approach appeared to be suited when structural macroeconometric models were built after the pioneer work of Jan Tinbergen, of Lawrence Klein and more generally of the Cowles Commission. Shocks no longer concerned just aggregate demand; in particular they were also present in the formation of prices and wages, or in the terms of foreign trade. But they were all viewed as "errors on the equations", each equation representing a behavioural law, an adjustment law or a market equilibrium law; these errors were commonly meant to be fairly small and to be well represented by simple stochastic processes, most often by purely random processes or, when more statistical sophistication spread out in the 1960s, by low-order autoregressive processes.

This errors-on-equations approach was conveyed by statistical teaching about linear regressions, first simple regressions, later multiple regressions. The transposition to systems of simultaneous equations, which was required by the needs of economists, did not question the nature of the basic stochastic assumptions; neither did it much discuss the linearity, which was most often maintained. These features will appear in the next part of the chapter, which will concern the simultaneous equations methodology.

Within this approach the observed variables were divided between two groups, exogenous ones which were not to be explained within the model or theory but had to occur in the explanation of variables of the second group, endogenous ones. Exogenous variables were assumed non-random, an assumption that was then considered convenient and appropriate for the contemplated analysis, whereas presence of random errors in equations implied that endogenous variables were random. The stochastic processes ruling these latter variables were not stationary because of the impact of changes in the exogenous variables. A stationary component could, however, be isolated, whose autocorrelation properties mainly reflected the role of the transmission mechanism described by the model.

When the use of structural macroeconometric models was attacked by strong criticism in the 1970s and 1980s, the stochastic approach had to be reconsidered. The main alternative econometric methodology, the VARs, conveyed a wholly different approach: instead of imputing the stochastic features to errors on equations, it stressed the vision of an overall multidimensional process embracing all observed variables, or rather deviations of those variables from their trends, even the vision of a stationary linear process.

The idea that the process could be the seat of shocks was accepted; but most often, for the identification of those shocks, each one of them was thought to bear on a variable rather than on an equation. We saw this approach at work in Section 5.6 of Chapter 8 dealing with monetary policy and we shall see it again in Part 4 within a more general context.

The VAR approach was not the sole innovation to have emerged during the last three decades and to be concerning us at this stage. Even if we limit attention to approaches which, like the VARs, wanted to "let the data speak" more freely than in the earlier structural modelling approach, we may recall that in Chapter 8, Section 7.4, we already introduced the concept of stochastic trends; we shall have more to say about this concept and its implications in a moment. The new vision according to which most economic trends would be random to some degree now contributes to shape our macroeconomic understanding.

But the most dramatic proposed revisions came with the claim, voiced by some economists, to build afresh macroeconomic theory, incorporating into it from the start a representation of the main types of uncertainties and of the formation of private expectations. This was advocating a stochastic approach to theoretical modelling, in a sense the opposite of the VAR approach whose target was empirical research.

To this theoretical ambition new classical macroeconomics added two assertions: all markets fully clear and expectations are rational, in the same way as more generally all private behaviours are rational. Initially the approach was not specific about the nature of uncertainties to which the economic system was exposed. Specific assumptions were made in this respect when particular models were built, such as the one examined in Section 5.2 of Chapter 8, which played an important part in discussions about monetary policy. But these models appeared so particular that it would have been unfair to say that their authors ruled out actual existence of other sources of uncertainty than those introduced.

Claiming generality came in the decade of the 1980s when the birth of real-business-cycle theory was announced[30]: the main uncertainties responsible for macroeconomic fluctuations were said to be real: they concerned the conditions of production and had to be introduced in production functions. Substituting supply shocks for demand shocks had quite significant macroeconomic implications, which we shall discuss.

The stochastic approach was also more fundamentally changed with respect to what it was with structural models. Real business cycle (RBC) modelling was not just introducing shocks as errors in different equations from what was common before: in the equation connecting output with its inputs, rather than

[30] The two founding articles were: F. Kydland and E. Prescott, Time to build and aggregate fluctuations, *Econometrica* **50** (1982) 1345–1370; J. Long and C. Plosser, Real business cycles, *Journal of Political Economy* **91** (1983) 39–69.

in the equations concerning the formation of demand. Inspired by new classical macroeconomics, it also meant that the stochastic process of the shocks had to be viewed as systematically affecting the behaviour of agents, which were assumed to rationally react to uncertainty; changes concerning this process, for instance, its autocorrelation features, then concerned household behaviour and the formation of prices and wages as well as the production function.

The real-business-cycle movement also brought a new methodology for dealing with econometric aspects. This methodology of data analysis, which will be studied here in Part 4, is such that reference to the stochastic properties of the shock processes becomes less binding. Accordingly the stochastic approach is more loose when data are confronted with theory.

As we shall see, the RBC methodology can accommodate almost any kind of departure from the fundamental hypotheses of new classical macroeconomics, market clearing and rational expectations. This is why we shall consider the RBC modelling several times in what follows, mainly in Section 2.3 devoted to the implications of the hypotheses of new classical macroeconomics, hence in particular to the models of the strict real-business-cycle school, but also more quickly in the following sections of this part, where other models studied by less faithful proponents of the approach will be alluded to.

2.1.7. *Trends, ARIMA processes and cointegration*

We must now briefly consider a few technical complements about the introduction of stochastic processes in the modelling of business fluctuations. Indeed, appropriate models may lead us beyond the domain of validity of stationary linear processes. Interference between trends and cyclical movements may require introduction of non-stationary processes. Asymmetry between expansions and depressions may reveal that non-linear processes are involved.

Discussing in Chapter 8, Section 7.4, the empirical analysis of shocks in macroeconomic time-series, we saw why the practice of combining stationary processes with deterministic trends was not everywhere appropriate. Focussing then on econometric analysis and forecasting we introduced the concept of difference-stationary stochastic trends. This led us to also entertain the consideration of higher-order difference-stationary processes and of cointegrated multidimensional processes. Finally we realized that, within a general stochastic approach to the study of time series, the distinction between permanent and transitory shocks was blurred.

We focus now on the overall visions underlying theoretical modelling rather than on econometric analysis geared towards direct application. We also think about full business cycles rather than simply about the joint dynamics of inflation and employment. Close connections between these alternative concerns, however, exist; in particular combination of trends with movements of shorter periodicities has to be accounted for in both cases (at present we have to remember that business cycles do not operate in a vacuum but rather in interaction with long-run growth). So, a common mathematical framework serves in both cases: for the development of the theoretical models discussed in this part of the chapter, as well as for the elaboration of econometric techniques, a subject approached in Section 7.4 of the preceding chapter and to be examined again later in this chapter.

Elaboration of, and familiarization with, the mathematical framework in question owe much to the needs of macroeconomic dynamics. This book is of course not the proper place for a complete presentation of the framework[31]. But a brief introduction, which will extend the mathematical digressions made earlier in the present chapter, will aim at bringing along the basic elements of the mathematical culture from which advanced econometric theory and macroeconomic modelling both draw.

The central concept is provided by *ARIMA processes* (autoregressive integrated moving-average processes). A unidimensional process $\{x_t\}$ is ARIMA if there is a white noise process $\{\varepsilon_t\}$, a number m, a positive integer d, two finite polymonials $A(L)$ and $B(L)$ of the lag operator L such that:

$$B(L)(1-L)^d x_t = m + A(L)\varepsilon_t \tag{65}$$

and if moreover all roots of the characteristic equation (14) corresponding to the polynomial $B(L)$ have modulus smaller than 1.

It is implicit in this definition that the first elements in the two polynomials are non-zero constants a_0 and b_0, which may be taken as equal to 1 without loss of generality. If the orders of the polynomials are respectively h for $B(L)$ and k for $A(L)$, the process is said to be ARIMA of order (h, d, k).

The qualifier "integrated" in the denomination of the process defined by (65), and still more the precision "integrated of order d", call to mind the fact that, after the difference operator $(1-L)$ has been applied once or d times, the process becomes stationary, even ARMA. Note also that "integrated of order d" means that the polynomial multiplying x_t in an equation such as (65), or rather the characteristic equation associated with this polynomial, has exactly d roots equal to 1.

Let us for a minute look at the kind of non-stationarity implied by Equation (65). With $d = 0$ the process $\{x_t\}$ would be stationary. In a sense the value

[31] For the purpose see C. Gouriéroux and A. Monfort, *Time Series and Dynamic Models* (Cambridge University Press, 1997). Definitions given below in our text will, at places and for simplicity, be a little more restrictive than needed.

of the integer d gives the force of non-stationarity; this may be seen from the deterministic unidimensional recurrence equation:

$$(1-L)^d x_t = m. \tag{66}$$

For $d=1$ its solution is:

$$x_t = x_0 + mt. \tag{67}$$

For $d=2$ it is:

$$x_t = x_0 + \left(x_0 - x_{-1} + \frac{m}{2}\right)t + \frac{m}{2}t^2. \tag{68}$$

So, if m is positive, the solution tends to grow like t, or like t^2, more generally like t^d. Also, the presence of initial values of x_t in all the terms of the solution, except the one with the highest power d in t, seems to be a sign that the introduction of random shocks will mean that those shocks will have permanent effects, which will be all the stronger as d will be higher. We may even rather speak of *persistent* effects, since they will grow with time if $d > 1$.

The intuition is correct. It can be given a rigorous formalization for unidimensional processes, and for multidimensional ones as well. In the latter case x_t, m and ε_t are n-vectors, $B(L)$ and $A(L)$ are square matrices of polynomials and the definition given above applies, the characteristic equation now concerning the determinant of $B(L)$. The process $\{x_t\}$ can be written as a sum of $d+1$ components y_t^r, where y_t^{d+1} is stationary and, for $r \leqslant d$:

$$(1-L)^{d+1-r} y_t^r = K_r \varepsilon_t \tag{69}$$

the matrices K_r being of course computed from $A(L)$ and $B(L)$. We may say that the first r components are explosive, all the more so as $d+1-r$ is.

Existence of one (if $d=1$) or several explosive components, acting as a stochastic trend or as components of such a trend, seriously complicates the estimation of periodicities of the fluctuations generated by the stationary component. If the order of integration d is known, a natural idea is to estimate the spectrum of the transformed series[32] $(1-L)^d x_t$, but this spectrum concerns the sum of all components similarly transformed. Retrieving an estimate of the spectrum of the untransformed stationary component will be required.

[32] We also speak of the *filtered series* after application of the filter $(1-L)^d$.

For instance, in the case of $d = 1$ and univariate analysis, if $f_1(\omega)$ and $f_s(\omega)$ are the spectral densities of respectively $(1 - L)x_t$ and of the stationary component of x_t, the relation between the two is given by:

$$f_1(\omega) = K_1 \frac{\sigma_\varepsilon^2}{2\pi} + 2(1 - \cos\omega) f_s(\omega). \tag{70}$$

In principle estimation of the first term of the right-hand member could follow from estimation of $f_1(\omega)$ at $\omega = 0$, because a ARMA process for the stationary component implies that $f_s(0)$ is finite. But in practice estimation of $f_1(0)$ usually appears very unprecise and indeed presence of a higher-order non-stationary component or of a deterministic trend, even with a very small weight, makes $f_1(0)$ infinite. We shall have opportunity to discuss the difficulty again in the next Section 2.2.

By definition in a multidimensional ARIMA process all variables, x_{it} say, defining the n-vector x_t have the same order of integration; hence they are explosive to the same degree. But this does not rule out that the evolutions of two or several of these variables might be strongly linked. In particular there may exist a linear combination:

$$\alpha' x_t = \sum_{i=1}^{n} \alpha_i x_{it} \tag{71}$$

that explodes less fast than each of the variables or perhaps does not explode at all, hence has a lower order of integration, perhaps is not even integrated. This is the case if the matrix K_1 appearing in (69) for the first component y_t^1 is singular. The process x_t and the variables x_{it} appearing in the linear combination (71) with non-zero coefficients are said to be *cointegrated*; the vector α is said to be a *cointegrating vector*. Once this concept has been introduced, many distinctions are possible: the order of integration of $\alpha' x_t$ may be $d - 1$ or less; there may be several linearly independent cointegrating vectors.

This concept of cointegration is closely related to that of *error correction*, often found in the specification of dynamic economic models. Indeed, in Chapter 8 we alluded to cointegration in Section 7.4 when dealing with the multidimensional generalization of the "difference-stationary stochastic trend". We then referred to the simultaneous evolutions of wages and prices and to Section 6.2 of the same chapter, where an error–correction term had been introduced in the modelling of the joint wage and price evolution, precisely in order to represent the link between the movements of the two variables. The relation between the two concepts of cointegration and error correction may be made mathematically rigorous[33].

2.1.8. *Non-linear processes*[34]

It will appear in what will follow that linear processes, although providing quite sufficient approximations for most purposes, cannot well account for all features of business cycles. For instance, asymmetries between depressions and expansions were often recognized in the past; recent tests, starting with those due to S. Neftci[35], confirmed the fact. It is of course possible to introduce non-linearities in the specifications, at the cost of more complex mathematical and econometric analysis. J. Hamilton has explored a natural way of dealing with the difficulty and presented a rather extensive treatment for the case when considering just a univariate stochastic process suffices. This treatment, although concerning a particular specification, gives a good insight into the differences between linear and non-linear analysis.

The main idea recognizes from the start that the dynamics differs in recessions from what it is in expansions. Therefore at the very least, a state variable s_t should appear in the formalization, a variable that will take either of two values 0 or 1, depending on whether period t belongs to a depression or an expansion. It will not be exactly a dummy variable because the analyst may not know for sure its value. So s_t will be a component of the stochastic vector whose process will be analysed.

The simplest assumption to be made about the process of s_t itself, taken in isolation, specifies that it is a first-order Markov process. Such is precisely what Hamilton assumes. There are two positive numbers p and q, respectively giving the probability that if $s_{t-1} = 1$ (or $s_{t-1} = 0$) also $s_t = 1$ (or $s_t = 0$). In other words the transition probabilities are:

$$\begin{aligned} &\text{Prob}\{s_t = 0/s_{t-1} = 1\} = 1 - p, \\ &\text{Prob}\{s_t = 1/s_{t-1} = 0\} = 1 - q \end{aligned} \tag{72}$$

and this holds independently of what may have been the earlier values $s_{t-\tau}$ for $\tau > 1$. Typically p and q are larger than $1/2$, especially if the unit time period

[33] R. Engle and C. Granger, Cointegration and error correction: representation, estimation and testing, *Econometrica* **55** (1987) 251–276.

[34] We shall here draw from the article by J. Hamilton, A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica* (March 1989).

[35] S. Neftci, Are economic time series asymmetric over the business cycle?, *Journal of Political Economy* **92** (1984) 307–328.

is a quarter; we also expect $p > q$ since recessions have been observed to be shorter than expansions.

From the Markov assumption we may draw a few consequences. First, if $\{s_t\}$ is stationary, as it has to be in the long run, the unconditional probability π that $s_t = 1$ may be computed from the conditional probabilities:

$$\pi = \pi p + (1 - \pi)(1 - q). \tag{73}$$

Hence

$$\pi = (1 - q)/\big[2 - (p + q)\big]. \tag{74}$$

Second, a kind of autoregression applies to $\{s_t\}$, namely:

$$s_t = (1 - q) + \lambda s_{t-1} + v_t \quad \text{with} \quad \lambda = p + q - 1. \tag{75}$$

The new random variable v_t is defined as follows: independently of what are $s_{t-\tau}$ for $\tau > 1$, v_t is equal to $1 - p$ with probability p and to $-p$ with probability $1 - p$ if $s_{t-1} = 1$, it is equal to $q - 1$ with probability q and to q with probability $1 - q$ if $s_{t-1} = 0$. A simple calculation shows that the conditional and unconditional expected values of v_t are equal to zero. It can moreover be proved that v_t is uncorrelated with $s_{t-\tau}$ for $\tau > 0$, but that it is not independent of $s_{t-\tau}$; in particular:

$$\begin{aligned} E\big[v_t^2/s_{t-1} = 1\big] &= p(1 - p), \\ E\big[v_t^2/s_{t-1} = 0\big] &= (1 - q). \end{aligned} \tag{76}$$

The unconditional variance $\sigma_v^2 = E(v_t^2)$ is equal to $\pi(1 - p)(p + q)$. The distinction between lack of correlation, which holds, and lack of dependence, which does not, is important for what follows. It means here that (75) should not be read as showing that $\{s_t\}$ is autoregressive: the process could just be said to be "second-order autoregressive".

The state variable s_t is not directly observed; we commonly infer its value from the evolution of real GDP. But this evolution is also subject to a trend and to erratic shocks, within expansions and recessions as well. Rules of thumb were adopted in order to tell us after how many quarters of decline we should be right in inferring that we entered a recession. Quite naturally econometricians concerned either about business diagnosis or about characterization of growth trends want to find more firmly grounded procedures.

In order to clarify the nature of the problem, Hamilton first states it in a simplified version. Assume that we observe a variable x_t that differs from s_t just by an independent ε_t the process of which is a Gaussian white noise with variance σ_ε^2:

$$x_t = s_t + \varepsilon_t. \tag{77}$$

Hamilton then shows that the process $\{x_t\}$ is ruled by an equation:

$$x_t = \lambda x_{t-1} + (1 - q) + u_t - \theta u_{t-1} \tag{78}$$

in which u_t is the sum of two independent second-order linear processes, defined one from $\{\varepsilon_t\}$ the other from $\{v_t\}$, both having zero unconditional expected value; the number θ can be computed from σ_v^2, σ_ε^2 and σ_u^2 (the unconditional variance of u_t); the correlation of u_t with any $u_{t-\tau}$ (for $\tau > 0$) is equal to zero. We may moreover assume $\{s_t\}$, hence $\{v_t\}$, to be stationary: this is an innocuous assumption since we are far from the beginning of times when the first-order Markov process started. So, the processes $\{u_t\}$ and $\{x_t\}$ are also stationary. This means that (78) gives, for $\{x_t\}$, the decomposition whose existence is established by Wold's theorem, the deterministic component being the constant $(1 - q)/(1 - \lambda)$ and the second-order linear component being $\{(1 - \lambda L)^{-1}(1 - \theta L)u_t\}$.

It would be tempting to use (78) for forecasting purpose, once λ, q and θ would have been closely estimated. For instance, having observed x_t and computed u_t from observed current and past values of x_t we would forecast x_{t+1} by:

$$x_{t+1}^* = \lambda x_t + (1 - q) - \theta u_t. \tag{79}$$

Although unbiased, x_{t+1}^* is, however, not an optimal forecast if the latter is defined as minimizing the expected square error of forecast, which requires:

$$x_{t+1}^e = E[x_{t+1}/x_t, x_{t-1}, \dots] = E[s_{t+1}/x_t, x_{t-1}, \dots] \tag{80}$$

(the second equality directly follows from (77) and the fact that ε_{t+1} is independent of all variables up to t). Now let:

$$\varphi_t = \mathrm{Prob}\{s_t = 1/x_t, x_{t-1}, \dots\} \tag{81}$$

a quantity which can be computed from observation once p, q and σ_ε^2 have been closely estimated. We derive from (80):

$$x_{t+1}^e = \pi + \lambda(\varphi_t - \pi). \tag{82}$$

This is in general a fairly different value from that given by (79).

This case well shows why non-linear analysis may be superior to linear analysis, to which we are led by the Wold's decomposition of the stationary process of interest. As a function of the series of observed values of x_t up to t, Equation (79) is linear since u_t is computed from this series by repeated application of the linear equation (78). On the contrary formulas worked out by Hamilton for the computation of φ_t are definitely non-linear. But they lead to superior forecasts, hence to a superior representation of the x_t process.

In order to model the evolution of real GDP, y_t say, Hamilton then proposes the following model:

$$y_t = \alpha_1 s_t + \alpha_0 + y_{t-1} + \eta_t, \tag{83}$$

where $\{\eta_t\}$ is a Gaussian autoregressive process of order r, which has zero mean and is independent of $\{s_t\}$. This means that, except for the effect of the random shocks generating η_t, y_t increases by the (probably negative) amount α_0 in recessions and by $\alpha_0 + \alpha_1$ in expansions. Hence the phrase "Markov trend in levels" to characterize the model (83). Hamilton also discusses a multiplicative model based on an equation otherwise similar to (83); the phrase "Markov trend in logs" characterizes this alternative model, which is used for an econometric fit of the quarterly series of the US real GNP from 1951 to 1984.

The maximum likelihood estimate concludes that the growth rate is equal to $\alpha_0 = -0.4\%$ per quarter during recessions and to $\alpha_0 + \alpha_1 = 1.2\%$ per quarter during expansions. The estimate also identifies the dates of shift from one state to another; they are quite close to those given by the *ad hoc* procedure worked out by the National Bureau of Economic Research.

Hamilton then discusses which assessments differ, in this particular application, depending on whether linear or non-linear analysis is used. He first shows that, if attention is focused on estimation of the autocorrelogram or on the expected values of the coefficients of a linear autoregression, then the two analyses lead to quite similar conclusions, even though the linear analysis was inspired by a wrong model. However, some predictions announced by the Markov model are not consistent with simple ARIMA representations but are validated on the case examined. In particular the ARIMA fit of log GNP is made significantly more consistent with the data if the estimate of φ_{t-1} is added to the right-hand member: knowing whether s_{t-1} is likely to have been 0 or 1 is an additional useful information beyond what is captured in the ARIMA regression. There is also a significant negative correlation between the ARIMA square residual and the estimate of φ_{t-1}, precisely the kind of heteroskedasticity announced by the Markov model.

2.2. Stylized facts

We want now to focus attention more systematically on the empirical knowledge of business cycle facts. This knowledge accumulated since long in the past and motivated attempts at explaining business cycles. It had to be reported somewhere in the present book. The empirical work of the last few decades having been substantial, the few next pages presenting such a report should be found to have here their natural location.

We shall of course not cover the full extent of the knowledge in question; it is too varied. But some results holding quite generally at the most aggregate level will be examined. The following sections will provide good opportunities for presenting additional significant facts about particular phenomena which will then be scrutinized.

Our main interest at present is in "stylized facts". We already introduced the concept in Chapter 4, Section 1.4 about the short-run fluctuations in the demand for labour. We shall not repeat what was written there, in particular about the risk of subjective biases in the selection of the facts so identified. The intention and the claim of authors of such facts are to provide descriptive indicators which were obtained independently of the consideration of business cycle theories and may serve for tests of those theories.

We know that some preconceptions are present behind the scene, for instance, in the selection of the magnitudes to be observed or in the definition of the aggregates. But the risk of distortions arising from such origins is felt to be weak, because the indicators are meant to agree with the widely accepted macroeconomic language.

The real risk rather stands in the opposite direction, namely that we learn too little from the descriptions so provided. Movements affecting economic activity are the combined result of so many factors, each one subject to particular ventures, that we should not expect that the same pattern always holds, except for a few main features which experience has shown to be most often apparent. Indeed, as soon as we reflect about the determinants of these movements and about their effects, as we did for instance in Sections 1.5 to 1.7, we realize that even qualitative conclusions may be hazardous.

For the short report to be now given we shall first consider the accumulated knowledge as it stood in the 1970s, then move on to what was learned more recently, not only because of new developments in economic activity, but also by application of a new method for the identification of business cycle facts.

1. In older times since the nineteenth century, the *empirical methodology* was not tightly defined, or rather it had a different nature. The main problem was then to collect sufficiently long and homogeneous series of data, on a sufficient number of variables, for being in a position to positively assess what overall movements had been: how did narratives, given by people who lived through those times, stand when they were compared with quantitative data? Economists are no longer now much concerned with this problem. The collection of comprehensive data, their adjustments so that they measure the same variables through time, their aggregation so that they are significant of overall activity or conditions, all such operations were taken over by statisticians, who apply methodological standards progressively defined, improved and extended to new areas. On the whole macroeconomists are happy with this distribution of roles, to such a degree that they often overlook that statistical series have not, and cannot have, perfect accuracy.

In former times each empirical research also had to adopt a methodology for the analysis of the available time series. People involved were exchanging experience about the matter, in particular on the occasion of meetings, so that their respective methodologies were inspired by similar concerns and had much in common. This is fortunate since we obviously cannot go here into a scrutiny of these old methodologies, except for giving a look at the practices developed in the country which was most advanced in this respect, the USA[36].

We shall even limit attention to methods used by a team working in New York at an office of the National Bureau of Economic Research, under the main leadership of W. Mitchell from 1920. These methods were explained and applied in a major book coauthored by A. Burns and W. Mitchell[37]. They are best viewed as taking place within a framework organizing the graphical scrutiny of time series and the characterization of their main features. They aim at fitting within the following statement about business cycles, a statement meant itself to be empirically based: "A cycle consists of expansions occurring at about the same time in many economic activities, followed by similarly general recessions, contractions, and revivals which merge into the expansion phase of

[36] For an independent reference see E. Wagemann, *Konjunkturlehre* (Berlin, 1928).

[37] A. Burns and W. Mitchell, *Measuring Business Cycles* (Columbia University Press, New York, 1946). For later development of the methods see A. Burns, *The Business Cycle in a Changing World* (Columbia University Press, New York, 1969).

the next cycle; this sequence of changes is recurrent but not periodic; in duration business cycles vary from more than one year to ten or twelve years; they are not divisible into shorter cycles of similar character with amplitudes approximating their own".

The methodology, which will be only sketched in this presentation, is iterative in the sense that what was achieved at one step may be refined at a later step. Search uses a reference chronology dating turning points (peak-trough-peak...). The initial chronology may be found in one or several series concerning global economic activity, or alternatively in qualitative evidence drawn from economic history. Available time series are then examined, one by one, against the reference chronology, after correction of the series for seasonal variations and, when it is manifest, for the trend.

Two main questions are then asked. First, what is specific in the pattern of the series under scrutiny? In particular, is it leading, coincident or lagging with respect to the reference cycle? Answers to the question are interesting in themselves in order to characterize how expansions and contractions diffuse. They also are the basis for the revision of the reference chronology: at each date are counted how many series reach a peak and how many a trough; the dates of the peaks and troughs of the overall cycle then appear.

2. Let us now take stock of *what was known in the 1970s about business cycle facts* and, for so doing, let us draw from an article by V. Zarnowitz who was long associated with the National Bureau of Economic Research[38]. First, let us quote him:

> "The term "business cycle" is a misnomer in so far as no unique periodicities are involved, but its wide acceptance reflects the recognition of important regularities of long standing. The observed fluctuations vary greatly in amplitude and scope as well as duration, yet they also have much in common. First, they are national, indeed often international in scope, showing up in a multitude of processes, not just in total output, employment and unemployment. Second, they are persistent – lasting, as a rule several years, i.e., long enough to permit the development of cumulative movements in the downward as well as in the upward direction. This is well established by the historical chronologies of business cycles in the United States, Great Britain, France and Germany, a product of a long series of studies by the National Bureau of Economic Research" (p. 525).

[38] V. Zarnowitz, Recent work on business cycles in historical perspective: a review of theories and evidence, *Journal of Economic Literature* (June 1985).

In the US the average duration of a cycle was 4 years between 1854 and 1933, 5 years between 1933 and 1982. Business cycles had tended to be longer in Europe than in the US. However, before World War II, more than 60 per cent of the cyclical turning points could be matched for all four countries covered by the NBER chronologies, and only 10 per cent could not be matched at all.

Private fixed investment expenditures, while much smaller in the aggregate than consumer spending, had much larger cycles in percentage terms. Inventory investment was also strongly procyclical. Business profits showed very high conformity and much greater amplitude of cyclical movements than wages and salaries, net interest and rental income. The level of industrial prices tended to have wider fluctuations than the levels of retail prices and wages.

Particular attention was brought in the US to leads and lags among major economic variables, and results did not seem special to that country. Of the various common components of aggregate demand, only change in business inventories and residential fixed investment appeared to be clear leading indicators, the others being roughly coincident with output, except business expenditures for new plant and equipment, which was lagging. But related data about decisions concerning, or likely to concern, demand components appeared leading, such as net business formation, contracts and orders for plant and equipment, rate of capacity utilization, index of consumer sentiment, new orders for consumer goods. About the labour market, the average workweek, the layoff rate and productivity were leading, long-term unemployment was lagging. Corporate profits and the ratio of price to unit labour cost were leading; unit labour cost and the labour share in national income were lagging. The growth rate of the quantity of money, change in consumer credit and private borrowing were leading, interest rates were lagging.

All these results seem to be fairly robust and can now be confirmed by the data concerning the two last decades. In some cases a long lag or lead amounts to counter-cyclicality, or to close to it. The remark concerns the mark-up at one extreme, the labour share at the other extreme, facts about which we shall have more to say later on. The reader is also advised to compare the above evidence with what was written earlier here in Sections 1.2 and 1.5 to 1.7.

3. As a by-product of the research initiated by the real-business-cycle movement, *a new methodology* emerged in the 1980s for the empirical analysis of business cycles. It argued particularly in favour of a kind of norm for the quantification of business cycle facts and it now provides a set of interesting results obtained and presented in conformity with this norm. Let us look first at the methodology, then at the set of results.

The statistics to be computed and considered derive from time series through the application of a linear analysis. This is why most series may be first transformed to the logarithms of their usual values. The next step consists in establishing "detrended" series: each original series is processed through a specific linear filter called the "Hodrick–Prescott filter". The operation amounts to decomposing the (log) series, x_t say, into a growth component x_t^g, the trend, and a cyclical component $x_t^c = x_t - x_t^g$; with the filter in question the decomposition minimizes the particular loss function:

$$\sum_{t=1}^{T} \left(x_t^c\right)^2 + \lambda \sum_{t=2}^{T-1} \left[\left(x_{t+1}^g - x_t^g\right) - \left(x_t^g - x_{t-1}^g\right)\right]^2, \tag{84}$$

where λ is a constant number chosen for the whole analysis. The higher is λ the closer is $\{x_t^g\}$ imposed to be to a linear trend. It is now conventional to take $\lambda = 1{,}600$ when quarterly series are processed. This has the effect that fluctuations with periodicity in the range of 3 to 5 years are most emphasized in the cyclical component, whereas periodicities longer than 8 years are almost fully removed[39]. Once detrended series are available, second order moments of the x_t^c components are computed and compared. Their variances are compared with that of y_t^c, the cyclical component of real GDP: some series are more volatile, others are less. The correlation coefficients between $x_{t-\theta}^c$ and y_t^c are computed, in some cases only for $\theta = 0$, in others for various leads ($\theta > 0$) and lags ($\theta < 0$) of x_t^c with respect to y_t^c. Such correlation coefficients may in principle be also computed between the cyclical components of different series.

Broadly speaking, the operations so performed and the statistics so obtained are appropriate for giving business cycle stylized facts. We earlier realized how significant for the study of business fluctuations were the relative amplitudes of movements exhibited by different macroeconomic variables, their correlation with fluctuations in output, as well as leads and lags between the movements in question. There is also a definite advantage for informed users to be well acquainted with the precise procedure through which stylized facts were obtained. However, the risk exists that the characterization be taken too much at face value, just because it becomes so standard. The set of summary statistics commonly provided does not exhaust, and cannot exhaust, the whole information which is contained in the original series and is relevant for the study of medium-run business fluctuations. The choice of the value of λ for detrending is not innocuous[40], not even the choice of limiting attention to series detrended

[39] See T. Cooley and E. Prescott, Economic growth and business cycles, in: T. Cooley (ed.), *Frontiers of Business Cycle Research* (Princeton University Press, 1995).

[40] Indeed some cycles have a long period, like eight years. Looking at the graph plotted by T. Cooley and E. Prescott (1995), op. cit., for the trend of the US real GNP (y_t^g computed as explained

by the Hodrick–Prescott filter[41]; variances are not perfect measures of volatility neither correlation coefficients perfect measures of association, as soon as we are not faced with a Gaussian world, outliers being too important or asymmetries being manifest; and so on.

4. *Evidence coming from application of the new methodology* has been accumulating during the decade of the 90s and may still be significantly enriched in the near future. Here is a short extract. D. Backus and P. Kehoe[42] analysed annual macroeconomic series of ten countries for which a period of about a century could be covered. The data base had its limitations, of which the authors were aware: annual data do not reveal most leads and lags, which are shorter than a year (indeed at the National Bureau most series were monthly); coverage of the statistical sources used for the compilation of the series much improved with time, which may lead not only to better accuracy but also to lower volatility for the most recent parts of the series[43].

However, a great deal of regularity is found in the cyclical behaviour of real quantities. Although the magnitude of output fluctuations varies across countries and over time (with the interwar period exhibiting everywhere large fluctuations), relations among variables are remarkably stable: investment is consistently 2 to 4 times as variable as output; both investment and consumption are strongly correlated with output; the trade balance is generally countercyclical, exhibiting larger deficits during booms than during depressions.

above), readers might think that a part of the cyclical movement was allocated to the trend. They may reach the same conclusion when looking at a similar graph plotted for the UK in R. Fiorito and T. Kollintzas (op. cit.).

[41] In an important article on the statistical properties of the filter, "Low frequency filtering and real business cycles" (*Journal of Economic Dynamics and Control* **77** (1993) 207–231), R. King and S. Rebelo exhibit the gain function of the filter as currently used and confirm that this filter allocates to the trend component the larger parts of the sine components that have periods of eight years and more. They then write: "Mechanical application of the HP filter as a detrending procedure removes important times series components that have traditionally been regarded as representing business cycle phenomena". They also show that, although better than the filter used for "exponential smoothing", the Hodrick–Prescott filter is unlikely to be optimal in the sense of minimizing the mean square error of extrapolations based on the representation it gives of the stochastic processes analysed. We may finally note that they draw attention to the dramatic alterations that HP filtering may introduce in measures of persistence, variability and comovements of macroeconomic time series. There is no surprise for logicians, given what we saw in Section 2.1, particularly in Subsection 7, but the fact shows that prefiltering is not at all an innocuous feature of the methodology at stakes.

[42] D. Backus and P. Kehoe, International evidence on the historical properties of business cycles, *American Economic Review* (September 1992).

[43] C. Romer, Is the stabilization of the postwar economy a figment of the data?, *American Economic Review* (June 1986).

Correlations between measured output movements in different countries are typically positive and more pronounced since 1950 than before the war. Also, price level changes have been in most countries more persistent since the last war than before. The correlation between the cyclical components of the price level and of output changed: up to the last war it was positive in most of the ten countries; it became consistently negative since then. An explanation for this change in the behaviour of the price level can probably be found in the importance of "supply shocks" in the 1970s and 1980s (these were really autonomous price shocks, which boosted inflation and depressed output); but this case well shows that stylized facts might be less robust about nominal variables than about real variables.

The literature now contains quite a few fairly convergent presentations of business cycle facts obtained by analysis of quarterly macroeconomic series of the last four decades. We shall select here the results of J.-P. Danthine and J. Donaldson[44]. They were obtained from data of eleven countries, the sample period running up to 1989 about, from a starting date in 1960 or earlier in most cases. The authors also quote interesting additional facts available in the literature.

Concerning real GNP and its main uses the results are similar to those found by Backus and Kehoe on long series: investment is again more variable than output by a factor of 2 to 4; both investment and consumption are strongly correlated with output; the trade balance is countercyclical; output fluctuations are positively correlated across countries[45]. Some additional studies give support to the conclusions derived by Zarnowitz from work earlier done at the National Bureau: they point to the part played by business inventory investment; they show that investment in household capital is highly procyclical and tends to lead movements in business investment.

Danthine and Donaldson closely look at the empirical association between the cyclical components of the GDP deflator and of output. The fact that the contemporaneous correlation is found systematically negative, in fluctuations during the three decades covered, indeed deserves attention, since it

[44] J.-P. Danthine and J. Donaldson, Methodological and empirical issues in real business cycle theory, *European Economic Review* (January 1993).

[45] The force of this intercountry output correlation is found puzzling by some economists because it is higher than that of analogous correlations for consumption and productivity (see D. Backus, P. Kehoe and F. Kydland, International business cycles: theory and evidence, in: T. Cooley (ed.) (1995) op. cit.). Could not an explanation of this fact come from the international transmission of demand shocks? Also relevant for subsequent discussion in the next section is the fact that productivity growth is found to be more correlated across manufacturing industries within one country than across countries within one industry (D. Costello, A cross-country, cross-industry comparison of productivity growth, *Journal of Political Economy* (April 1993)).

runs counter to what was often earlier asserted in the business cycle literature. Avoiding to explain the fact by the importance of special price shocks during the period, which was alleged here a moment ago, they want to stick to empirical evidence given by the methodology used; looking at lagged correlation coefficients[46] they find that the negative association dissipates when the price level is correlated with output a year and more earlier and that then a positive association even appears in half of the countries[47].

Considering the quantity of money and interest rates, few strong regularities can be detected: the monetary aggregates M_1 and M_2 tend to be weakly correlated with output[48]; long term rates are less variable than short term ones; only the latter tend to be systematically correlated with output, with a positive but weak correlation (so, during the period covered, monetary policy appeared mildly or not uniformly countercyclical, if it must be gauged by the nominal short rate)[49]. These results should be compared with those given in Chapter 8, Section 5.6; interpreted as stylized facts, the results of the VAR analysis may appear more significant than those now reported.

Some labour market stylized facts were already examined in Section 1.4 of Chapter 4: employment is lagging with respect to output, and somewhat less volatile; labour productivity is about as volatile as employment; hours per worker tend to be a leading indicator for output. Danthine and Donaldson confirm these facts on a larger set of countries and add to then results about the real wage, which is positively correlated with employment (except in one country), but appears to be procyclical in only five countries (acyclical in four, countercyclical in two)[50]. They conclude: "All in all, it appears that labor market

[46] This is the only case in which the authors report such lagged correlations. We may regret that no more attention is given to them in the article. For the US economy they are displayed in T. Cooley and E. Prescott (1995), op. cit. For the G7 countries a number of them are reported in the article by R. Fiorito and T. Kollintzas, which was examined in Section 1.4 of Chapter 4.

[47] Essentially the same phenomenon is detected when it is observed that fluctuations in the rate of inflation are positively correlated with fluctuations in output.

[48] However, if lagged correlations are considered, monetary aggregates of a quarter appear to be more strongly correlated with output one, two or three quarters later.

[49] However, R. Fiorito and T. Kollintzas (op. cit.) exhibit a quite systematic pattern in the association between the cyclical movements of the real long-term interest rate and of output: there is a negative correlation between this rate and output one year later, there is a positive correlation between output and the real rate one year later.

[50] The procyclicality of real wages had already been the object of some econometric debate. Indeed, in his footnote 26, V. Zarnowitz (op. cit.) gives reference to conflicting results reached by various econometricians. After the discussion in Subsection 1.5.2 we understand why it may be so. Procyclicality tends to be supported for the US by evidence coming in particular from microeconomic data (see M. Bils, Real wages over the business cycle: evidence from panel data, *Journal of Political Economy* (August 1985)). But the US is one of the countries in which Danthine

behavior is substantially different across countries, most likely reflecting cultural and institutional arrangements". Let us remember, however, that labour productivity was systematically found to be procyclical, its fluctuations somewhat leading those of output[51].

5. So is a broad account of the results obtained about stylized facts by application of the newly codified methodology. The reader is advised to reflect on the nature of what was so learned. Perhaps he or she will wonder: did we so grasp the full picture of business cycle facts, or just a partial view? Is there no more to remember than relative volatilities and correlations lagged by a few quarters? Is the concept of periodicity irrelevant for a description of what we call cycles?

It is refreshing to read an article which, although written by authors who were accepted as members of the real-business-cycle movement, do not stick to the standard procedure for detection of business cycle facts[52]. Analysing US quarterly data for the period 1950–1990, R. King and M. Watson first exhibit estimates a few spectral densities. They show in particular the importance of periodicities in the range from 5 to 10 years for the growth rate of output and of a wide range of real macroeconomic variables: there is actually a clear peak in this range, like in Figure 3, the density being definitely low for the small frequencies corresponding to the trend[53]. A similar peak also exists for the growth rate of the quantity of money (with a period of about 8 years) and for the rate of inflation (with a period of 13 years), but the peak less strongly dominates over the spectral densities of the trend than it does for the growth rate of output.

and Donaldson find procyclicality. Let us also note that results given by Fiorito and Kollintzas about the manufacturing sector broadly agree with those of Danthine and Donaldson. The reader may finally also refer to P. Geary and J. Kennan, The employment-real wage relationship: an international study, *Journal of Political Economy* (August 1982).

[51] A serious limitation of labour market facts commonly reported by user of the new methodology results from the castration imposing to ignore data on unemployment. However, M. Merz, Search in the labor market and the real business cycle, *Journal of Monetary Economics* (November 1995), breaks the rule when reporting some facts about the US economy, in particular that unemployment has high degrees of volatility and persistence, and that it is strongly negatively correlated with job vacancies.

[52] R. King and M. Watson, Money, prices, interest rates and the business cycle, *Review of Economics and Statistics* (February 1996).

[53] We note in passing that estimates of spectral densities are not very accurate, unless fluctuations follow a fairly simple and stable pattern and the series analysed are long with respect to the relevant periods. See, for instance, Section 7, Chapter 11, in E. Malinvaud, *Statistical Methods of Econometrics*, 3rd ed. (North-Holland, Amsterdam, 1980).

Differences in these observed periodicities are worth remembering, because they are at variance with the common notion of cyclical fluctuations stirring all macroeconomic variables. In particular the longer periodicities of the rate of inflation call to attention the fact that the medium- and long-run dynamics of inflation happens to have features of its own. The fact is not so surprising to us because we recognized the possible importance of exogenous influences on the inflation process (the variables z in the main formulation used in Chapter 8).

King and Watson use the spectral decomposition of each series for defining a cyclical component, which cumulates what comes out of the decomposition for the frequencies in the range corresponding to 6 to 32 quarters. They can then graph the time profiles of the cyclical components of a few variables. We then perfectly see what happened about the comovements of prices and output during the postwar period: between 1972 and 1984 two important waves occurred for both prices and output, output declining when prices were rising and vice versa. If we remove those twelve years the price series exhibits only small variations, but those are cyclical.

Finally King and Watson compute, over the whole sample, the volatilities, correlation coefficients and lagged correlations with output for the cyclical components of a number of their variables. They can then reach conclusions bearing on the facts commonly considered by the standard methodology, in particular the two following ones: the quantity of money is highly correlated with output; nominal interest rates are strongly negatively correlated with future output ($\text{cor}(r_t, y_{t+6}) = -0.74$); real interest rates move countercyclically and also lead output.

2.3. Price flexibility and rational expectations: the theory of real business cycles

2.3.1. *Introduction*

After the preliminaries exposed in the foregoing two sections, we now enter into the study of developments which, during the last two or three decades, bore on business cycle theory in various ways. We begin with developments originally inspired by the vision that is most at variance with the Keynesian one, studied in the first part of the chapter. According to this alternative vision business cycles would be due to the natural reaction of our economies to the technological shocks to which they are exposed. The phrase "real-business-cycle theory" will be used here in reference to this vision.

Interest in it was conveyed not only by the supposed realism of its substance but also, first, by dogmatic claims on what was alleged to be the correct scientific approach to macroeconomics and, second, by the diffusion of a new strategy for business cycle research, with its deductive part based on a new kind of models and its inductive part applying a new practice of empirical validation. In their most extreme form the claims in question stated that:

(i) full price flexibility and market clearing had to be assumed;
(ii) expectations had to be rationally formed by agents;
(iii) attention should focus on macroeconomic evolutions leading to efficient results.

There was so much in the real-business-cycle theory, strictly defined, that it was possible to accept part of it and to neglect the rest. In particular, a number of economists were attracted by the research strategy and applied it without accepting either the full real-business-cycle vision or the dogmatic premises listed above. We cannot say that a new school was so born, because there are disagreements on matters of macroeconomic substance within the group of those applying the research strategy in question. Here we shall speak of the "real-business-cycle movement", or more conveniently of the RBC movement.

After this Section 2.3 devoted to the real-business-cycle theory narrowly defined, the next one (2.4) will concern the work of the RBC movement. It will lead us up to a point where even a loose reference to the label RBC will progressively lose meaning. So, for the further study of related works and for reflections about the perspectives offered by the same strand of research we shall in Section 2.5 alternatively speak of the RBC approach or, probably better, of "dynamic stochastic general equilibrium". In any case the general equilibrium in question will take various forms, more or less substantially distant from the canonical concept introduced by the initiators of the movement.

The focus will shift back in Section 2.6 where perfect market clearing and rational expectations will again both be assumed, but in a different mood. Indeed, some economists worked on a research project that can be seen as a reaction to early writings of promoters of the real-business-cycle theory, and more generally to the teachings of the rational expectation school (see also Chapter 8, Sections 5.2 and 5.3). They wanted to explore the dynamic implications of premises (i) and (ii) above (market clearing and rational expectations) within less simple-minded environments than was most often done by the school in question. Some exhibited evolutions subject to quite different cycles from those which will be first studied below; they sometimes spoke of "endogenous cycles".

Such is the broad spectrum that will keep us busy during four sections. In this one we shall, first, consider a simple model, which will introduce us both to the real-business-cycle theory and to the research strategy of the RBC movement. We shall, second, examine the reason why the strict real-business-cycle vision has to be rejected if it is meant to be appropriate for analysis of the whole phenomenon of business fluctuations.

2.3.2. *A simple model of a stochastic dynamic economy*

The initial idea was to use for a new purpose the logical structure that had been earlier considered appropriate by some economists for studying the determinants of the long-run saving rate (see Chapter 6, Section 4.2, "*Market implementation of optimal growth*"). This purpose would now be to analyse the reactions of a competitive economy to recurrent shocks affecting the productivity of production, an economy in which prices and wages would be fully flexible so that all markets would always clear. The intention was to show that the model so conceived could reproduce many of the observed business cycle stylized facts. Let us then look without further ado at a simple model which captures the essential features of the first generation of models discussed by the RBC movement[54].

Conditions of production are specified in the same way as in Section 4.2 of Chapter 6 by what were then Equations (65):

$$y_t = u_t L_t \varphi(k_{t-1}) + \delta K_{t-1}, \quad k_{t-1} = \frac{K_{t-1}}{u_t L_t}. \tag{85}$$

Gross output y_t during period t is obtained from labour L_t and from capital K_{t-1} chosen at the beginning of the period from the stock of goods left at the end of period $t-1$; depreciation occurs at the rate δ. Labour augmenting technical progress is exogenously given and represented by changes in the factor u_t, the "technological level".

The new feature is now that $\{u_t\}$ follows a stochastic process: u_t is observed at the beginning of period t when decisions are taken about the period; but future values $u_{t+\tau}$ for $\tau > 0$ are not yet known. Given our recursive representation of production, ignorance about the future does not matter for the production sector.

Markets are open at the beginning of period t; they concern: (i) the services of labour, for which a (real) wage rate w_t, to be paid at the end of the period, is determined, (ii) loans of the good for the period, for which a (real) interest rate ρ_t is determined, and (iii) exchanges of the good to be delivered at the end of the period, the price being the "numeraire". The representative firm recruits L_t

[54] With respect to the model presented by T. Cooley and E. Prescott (1995, op. cit.) in their introduction to the book entitled *Frontiers of Business Cycle Research*, there will be two simplifications: the labour supply will be here exogenous and, in order to easily reach conclusions, we shall give a particular form to the production and utility functions.

for the period, borrows the quantity K_{t-1}, finally sells y_t and buys δK_{t-1} (to be both delivered at the end of the period). Given these competitive markets, equilibrium of the representative firm requires:

$$w_t = u_t\big[\varphi(k_{t-1}) - k_{t-1}\varphi'(k_{t-1})\big], \qquad \rho_t = \varphi'(k_{t-1}). \tag{86}$$

Like as in Chapter 6, the crucial part is the model of the representative individual, viewed at the initial date 0 and supposed to plan his activity for an infinite future[55]. This activity will consist in supplying the amount of labour L_t in period t and buying the amount of good C_t at the beginning of period t for consumption at the end of the period. The representative individual in this economy faces a very difficult problem because he must form expectations over the future price system in order to choose a strategy, which will determine his supplies and demands recursively when the value of u_t will be known and the markets will be open (see T. Cooley and E. Prescott (1995, op. cit.) bottom of page 13). The budget constraint will be the same as in Chapter 6, namely:

$$\sum_{t=1}^{\infty} \beta_t C_t = \sum_{t=1}^{\infty} \beta_t w_t L_t + K_0, \tag{87}$$

where K_0 is both the initial capital available for production in period 1 and the initial asset of the individual, while β_t are the market discount factor defined as usual by:

$$\beta_0 = 1, \qquad \beta_t = \frac{\beta_{t-1}}{1+\rho_t}. \tag{88}$$

For simplicity we assume that labour supplies L_t in all periods are exogenous. But the individual is free to choose any series of consumptions C_t he prefers, subject only to the constraint (87). For the preference function we make the same kind of assumption as in Chapter 6, but must now recognize that future consumptions are random variables. So the following mathematical expectation is chosen:

$$U = E\sum_{t=1}^{\infty} \gamma^t U(C_t). \tag{89}$$

[55] The RBC literature does not speak of a representative individual, but of a large number of perfectly identical individuals living for ever. Will this rhetoric make the model still more attractive?

We are familiar with the mathematical problem posed by maximization of this preference function subject to a linear budget constraint, here (87). We already faced this problem when dealing with consumer behaviour in Chapter 2, Section 1.8. The consumer was there planning up to a finite horizon, whereas our preference function now concerns an indefinite future. But we discussed the implications of this new feature both in Section 3.4 of Chapter 5 and in Section 4.2 of Chapter 6, when speaking of a transversality condition. For our present purpose the main point is to focus on the first-order marginal conditions, the "Euler equations", which take here the following form:

$$U'(C_t) = \gamma E_t(1 + \rho_{t+1})U'(C_{t+1}), \quad t = 1, 2, \ldots, \tag{90}$$

where E_t is the mathematical expectation operator, conditional on what is known at time t when the demand for consumption is expressed. Equations (90) and (87) would leave one degree of freedom for the determination of the consumer strategy, given the stochastic process of the ρ_t, or equivalently the process of the β_t. The transversality condition requires selection of the strategy maximizing C_1 among all those which will meet the budget constraint and the Euler equations, and will moreover be indefinitely sustainable ($C_t > 0$ for all t).

We want to know what the model implies for the stochastic process of its endogenous variables $\{y_t, C_t, K_t, w_t, \rho_t\}$. This will, of course, depend on the precise specification of its exogenous elements: the process $\{u_t\}$, the series $\{L_t\}$, the functions $\varphi(k)$ and $U(C)$, the parameters K_0, δ and γ. Mathematical derivation of the results will be difficult in almost all cases. This is why the RBC movement developed computational techniques for the needed solutions (see Chapter 2, by G. Hansen and E. Prescott, and Chapter 3, by J.-P. Danthine and J. Donaldson, in T. Cooley (1995), op. cit.). Since entering into a presentation of these techniques is out of question here, given the scope and purpose of this book, we shall consider a mathematically convenient specification, thanks to which we shall be able to grasp the whole logical chain from beginning to end.

This convenient specification is "the log-linear economy" already introduced in the second part of Section 3.4 of Chapter 5[56]. It concerns the functions $\varphi(k)$ and $U(C)$, more precisely here:

$$\varphi(k) = Bk^{1-\alpha} - k, \qquad U(C) = \log C. \tag{91}$$

[56] The log-linear economy was also introduced for a similar model with several industries in R. Radner, Optimal growth in a linear-logarithmic economy, *International Economic Review* (January 1966). That multi-sectoral model was used in J. Long and C. Plosser, Real business cycles, *Journal of Political Economy* **91** (1983) 39–69.

The second in Equations (86) can then be written as:

$$1+\rho_t=(1-\alpha)\frac{S_t}{K_{t-1}}, \tag{92}$$

where:

$$S_t=y_t+(1-\delta)K_{t-1}=C_t+K_t \tag{93}$$

(the second equation in (93) gives the equilibrium of the market for goods). The Euler equation (90) so takes the form:

$$\frac{K_t}{C_t}=(1-\alpha)\gamma E_t\frac{S_{t+1}}{C_{t+1}}. \tag{94}$$

Replacing S_{t+1} by $C_{t+1}+K_{t+1}$ and introducing:

$$h=\frac{(1-\alpha)\gamma}{1-(1-\alpha)\gamma} \tag{95}$$

we write (94) as:

$$\frac{K_t}{C_t}-h=\frac{h}{1+h}E_t\left[\frac{K_{t+1}}{C_{t+1}}-h\right]. \tag{96}$$

Considered for all values of $t=1,2,\ldots$, the system of these equations has an obvious solution:

$$K_t=hC_t \tag{97}$$

or equivalently given (93)

$$C_t=\frac{1}{1+h}S_t, \qquad K_t=\frac{h}{1+h}S_t. \tag{98}$$

This sharing rule for the division of the stock of goods that will be available at the end of period t is indefinitely sustainable. It can be shown that other sharing rules, fulfilling (96) and implying only information available at the beginning of period t, either are not indefinitely sustainable, at least under admissible specifications of $\{u_t\}$, or lead to lower values of C_1 (and also to lower values of the subsequent C_t). Thus, the optimal strategy implies the sharing rule (98).

Calculation is then easy: at each step t, given K_{t-1}, L_t and u_t, y_t, S_t, C_t, K_t, w_t and ρ_t are obtained from the above equations.

We are then in a position to study the comovements of these endogenous variables as a function of the series $\{L_t\}$, the stochastic process $\{u_t\}$ and the parameters. Let us consider the case when the underlying growth is assumed exponential, with the growth factor $g = \mu\nu$ and:

$$\log L_t = t \log \mu, \qquad \log u_t = t \log \nu + \varepsilon_t, \tag{99}$$

where $\{\varepsilon_t\}$ is a stochastic process with a zero mean value. Then the production function gives:

$$\log S_t = (1-\alpha)\log K_{t-1} + \alpha t \log g + \log B + \alpha\varepsilon_t. \tag{100}$$

Taking the sharing rule (98) into account, we find:

$$\log S_t = (1-\alpha)\log S_{t-1} + \alpha t \log g + \alpha b + \alpha\varepsilon_t, \tag{101}$$

where

$$\begin{aligned} \alpha b &= \log B + (1-\alpha)\log \gamma(1-\alpha), \\ \log S_0 &= \log\big[K_0/\gamma(1-\alpha)\big]. \end{aligned} \tag{102}$$

Introduce now the sure series $\overline{S}_t$ and the new variable s_t defined by:

$$\log \overline{S}_t = \left[t - \frac{1-\alpha}{\alpha}\right]\log g + b, \qquad s_t = \log S_t - \log \overline{S}_t. \tag{103}$$

In the equilibrium growth, this last variable follows a stochastic process which is simply derived from $\{\varepsilon_t\}$ by:

$$s_t = (1-\alpha)s_{t-1} + \alpha\varepsilon_t, \tag{104}$$

$$s_0 = \log K_0 - \frac{1}{\alpha}\big[\log B\gamma(1-\alpha) - (1-\alpha)\log g\big]. \tag{105}$$

If $\{\varepsilon_t\}$ was a pure white noise, $\{s_t\}$ would be a first order autoregressive process, which would not be stationary because of the effect of initial capital K_0. More generally, if $\{\varepsilon_t\}$ is a stationary process, $\{s_t\}$ is the sum of a stationary process and a transitory component which tends to zero with time. The stationary component is a smoothed image of $\{\varepsilon_t\}$ under a moving average of previous values with exponentially decreasing weights.

Because of the sharing rule the variables C_t, K_t and S_t are synchronic and have precisely the same degree of variability. The commonly studied investment $I_t = h[S_t - (1-\delta)S_{t-1}]/(1+h)$ and output $y_t = S_t - \omega S_{t-1}$, with $\omega = h(1-\delta)/(1+h)$, are leading and have a higher degree of variability, I_t more so than y_t in both respects because $\omega < 1-\delta$. Let us note in passing that this is precisely what was obtained with the multiplier-accelerator model of Section 1.2. Equation (92) shows that the real interest rate ρ_t is also leading with respect to consumption, just a bit more than output does. Since $w_t L_t + (\rho_t + \delta)K_{t-1} = y_t$, the real wage rate should then be lagging with respect to output. However, movements of all series C_t, I_t, y_t, ρ_t, ω_t are strongly correlated.

The results so found cannot claim to be representative, since they were obtained thanks to a convenient but unusual specification. In general the outcome will depend on the values given to the exogenous functions and parameters, particularly on how vary elasticities characterizing intertemporal consumption choices (the ratio $\gamma U'(C_{t+1})/U'(C_t)$ in the Euler equation) and capital-labour substitution in production (the factor $1+\rho_{t+1}$ in the Euler equation).

T. Cooley and E. Prescott (1995, op. cit.) give results obtained from a specification which better represents those used as a support for the real-business-cycle theory. There are two main differences from our specification: first, the production function is a true Cobb–Douglas, with a constant elasticity of substitution equal to 1, whereas our convenient production function has a smaller and varying elasticity of substitution; second, the supply of labour is endogenous. The logarithmic objective function of the representative individual contains both consumption C_t, as above, and leisure: the excess of a maximum feasible labour supply over the supply L_t freely chosen by the individual, in the same way as he chooses his demand for consumption C_t, that is, only subject to the intertemporal budget constraint (87).

This second main difference is, of course, fundamental for the implication of the theory about the fluctuations in employment. Since the real wage rate is procyclical, in the canonical model presented by Cooley and Prescott, in the model discussed above and in many others, the representative individual has the opportunity to substitute labour, when the wage rate is high, for consumption and leisure when it is low. It is then conceivable to find that the model, which assumes labour market clearing, predicts employment to be procyclical.

As we saw when studying labour supply in Chapter 2, reactions induced by changes in the real wage rate involve two effects going in opposite directions, a substitution effect, which taken alone would make the labour supply procyclical, and an income effect, which would make it contracyclical. The canonical model implies procyclicality; the substitution effect dominates because the intertemporal elasticity of substitution is high in the specification that was chosen, and because the representative individual perceives only a small impact of transient high or low wage rates on its permanent wealth, the right-hand member of Equation (87).

Endogenous fluctuations in employment make the main difference between the results derived by Cooley–Prescott from their model and results found here from our mathematically more convenient and simpler model. Also noticeable is, however, the fact that significant leads or lags hardly appear in their results.

2.3.3. *The theory of real business cycles*

That specific theory is now going to be easily explained and discussed. We shall begin with the arguments put forward by proponents of the theory, before we consider why the author of this book, among many others, is not persuaded by the arguments in question and has also some counter-arguments to provide[57]. For simplicity and convenience, quotes and references will be given, but they should not be interpreted as fully characterizing the views of the respective authors: these views have been in all cases more elaborated, with nuances which are not reported here.

Narrowly defined, so as to be correctly expressed by its label, real-business-cycle theory asserts that technological shocks dominate as the most important cause of business fluctuations and that effects of these shocks are well described by the neoclassical model of capital accumulation, which was earlier interpreted as a model of economic growth. The shocks are sufficiently recurrent to be represented by a stochastic process. So, the neoclassical model can be reformulated along lines sketched above. The theory claims to be validated by confrontation of its results with business cycle stylized facts.

The argument put forward in the first place claims that, as opposed to Keynesian dynamics, the theory now discussed has explicit and firm choice theoretic foundations borrowed from the well established microeconomic theory of the competitive equilibrium (see C. Plosser, op. cit.). Invoking the authority of G. Debreu seems to be meant as giving a sufficient guarantee. Based on such foundations, the theory of real business cycles draws the consequence of the sound idea that "in order to understand business cycles, it is important and necessary to understand the characteristics of a perfectly working dynamic

[57] The same pattern is present in the symposium published in the *Journal of Economic Perspectives* (Summer 1989), with the advocate C. Plosser, "Understanding real business cycles", and the opponent G. Mankiw, "Real business cycles: a new Keynesian perspective".

economic system" (also C. Plosser, op. cit.). Such was the alleged recommendation of J. Hicks.

Why are technological shocks important? Well, in his famous 1957 article[58] on the empirical analysis of US economic growth, R. Solow[59] showed that "technical change, in addition to the capital per worker, was an important source of variation in output per capita" (same reference as above). Since there are definite advantages in "an integrated approach to the theory of growth and fluctuations" it is natural to consider that the "Solow residuals", computed as explained in the 1957 article (see here Section 7.3 of Chapter 4), measure not only the long-run effect of technical progress but also short-run variations in the pace of this progress.

In addition, models built as was explained at the beginning of this section and fed with the series of Solow residuals viewed as the realization of the stochastic process of shifts to the production function well mimic the observed business cycle facts. "The simple model appears to replicate a significant portion of the behavior of the economy during recessions as well as other periods" (C. Plosser, op. cit.). "The match between the model economy and the observed data for the US economy is pretty good but clearly not perfect... The broad features of the model economy suggest that it makes sense to think of fluctuations as caused by shocks to productivity" (T. Cooley and E. Prescott, op. cit.).

For clarity let us be a little less vague already at this stage about the method used for the empirical validation so offered (we shall discuss the method more precisely in Section 4.5 of this chapter). The theoretical model is first "calibrated" in such a way that empirical equivalents to its abstract variables are defined, values are chosen for its parameters and a realization of the shock process is found (for instance, a series of observed Solow residuals). Numerical solutions of the calibrated model are computed, giving series of the variables. The Hodrick–Prescott filter is applied. Variances, covariances, auto- and cross-correlations are computed on the filtered series. The results are compared to the stylized facts obtained according to the new methodology explained in Section 2.2.

Doubts about the justifications given by proponents of the theory of real business cycles concern the three above categories of argument: the claim that the theory had firm microeconomic foundations, the strategic role to be recognized to irregularities in technical progress, the claim that the theory provided a good match to observed macroeconomic time series. As for the first

[58] Reference given on p. 488 of Volume A of this book.

[59] I might dream of a dissertation in the history of economic theory in which the likelihood of adhesion of Debreu, Hicks and Solow to the later theory of real business cycles would be tested.

claim, one might quibble about the robustness of the foundations given to that theory by the microeconomic models of the competitive general equilibrium: under which realistic conditions would such an extreme aggregation as the one implied by Equations (85) to (90) be justified? But there is a much more important objection: can we seriously speak of firm microeconomic foundations of a business cycle theory when that theory assumes permanent market clearing and when there is so much microeconomic evidence that changes in the degree of deviation from market clearing along business fluctuations are dominant features of these fluctuations, as perceived in particular by business men and workers? I shall not repeat here what was explained in Section 3.4 of Chapter 7; but I do not see how the evidence there listed can be dismissed.

As for the second claim and the central role attributed to irregularities in technical progress, it can only look strange to a microeconomist specialized in the study of technologies, from their first appearance, to their diffusion and in most cases to their replacement by new ones: this is a slow process operating in terms of years, often decades, not quarters[60]. Indeed, Solow residuals were not earlier interpreted as evidence of exogenous technological disturbances. "The standard explanation of cyclical productivity is that it reflects labor hoarding and other "off the production function" behavior. Productivity appears to fall in a recession because firms keep unnecessary and underutilized labor. In a boom the hoarded laborers begins to put out greater effort; output increases without a large increase in measured labor input" (G. Mankiw, op. cit.). The phenomenon was explained as a direct consequence of adjustment costs when we studied the short-term fluctuations in the demand for labour (Section 1.4 of Chapter 4). A number of macroeconomists also argued that productivity-improving technological changes were industry specific and that analysis of disaggregated data showed a high correlation, across industries, between short-run changes in productivity: the conjunction of these two facts appeared to make implausible the interpretation attributing observed irregularities in productivity movements to exogenous technological shocks[61].

[60] In many calibrations, particularly in the one presented for the prototype model by Cooley and Prescott, the technological shock process is chosen to be first-order autoregressive with an autocorrelation coefficient equal to 0.95. This would mean just a correlation of 0.36 between the shock in a quarter and the shock five years later. (Section 1.4 of Chapter 4.)

[61] Econometric investigation of this implausibility are more or less sophisticated. We shall mention one in Subsection 2.5.5. At this stage let us simply note that D. Costello, A cross-country, cross-industry comparison of productivity growth, *Journal of Political Economy* (April 1993), studied movements in productivity during the years 1960–85 for five major industries in six industrial countries. Her main conclusion is that short-run productivity growth is similar across industries in a nation, but less similar across countries in any specific industry. This seems to suggest that aggregate demand shocks, which have well known national features, are more important in ex-

Concerning finally the third claim, we shall not proceed now to a full discussion of the force of the argument stating the good match between the fluctuations generated by the model economy here at stakes and the aggregate time series describing the fluctuations observed in actual economies. This discussion is complex because it requires a serious examination of alternative inductive procedures for the test of theories, a subject that is delayed to Part 4 of this chapter. We may indeed remark that introduction to the empirical procedures used by the RBC movement, from the determination of stylized facts, to the calibration of the model and to the confrontation of the results drawn from the model with the facts, take a large part in the introductory chapter of *Frontiers of Business Cycle Research* (T. Cooley and E. Prescott, op. cit.). The main point to be made at this stage is that the tests provided, "the good match", has low power, in the sense that other theories could, just about as well, pass the same test. We already mentioned in passing that even the simple multiplier-accelerator model of our Section 1.2 was fairly well matching a number of facts; the main discrepancy, that investment was leading with respect to output, could easily be corrected by introduction of longer times to build capital.

2.4. The RBC movement

Advancement in the research programme of the RBC movement is presented in the book published in 1995 and edited by T. Cooley, the contributions brought over a decade by a large number of articles being individually identified. This will now be our main reference[62]. "The intent of [the] book is to describe the methods and problems of modern business cycle research, using the neoclassical growth framework to study the economic fluctuations associated with the business cycle"[63]. The preface of the book dates the emergence of the approach from the publication of two articles commonly quoted in this respect[64]. It then

plaining short-run productivity growth than are for the same purpose technological shocks, which ought in most cases to be industry specific.

[62] For convenience the reference is repeated here: T. Cooley (ed.), *Frontiers of Business Cycle Research* (Princeton University Press, 1995).

[63] First sentence of the introduction. The reader may note an ambiguity in the interpretation of the second part of the sentence. Does it restrict the scope of the book to only part of "modern business cycle research", or does it imply that all the research in question is using the neoclassical growth framework? The same ambiguity is present at many places in the book.

[64] F. Kydland and E. Prescott, Time to build and aggregate fluctuations, *Econometrica* **50** (1982) 1345–1370; J. Long and C. Plosser, Real business cycles, *Journal of Political Economy* **91** (1983) 39–69.

goes on writing: “The most important aspect of the real business cycle development is that it established a prototype and a set of tools for carrying the equilibrium approach forward. It combines general equilibrium theory with a set of tools for computing the equilibria of artificial economies and studying their empirical properties”.

For a brief survey of what was achieved so far we shall successively consider four alternative main objectives that various contributions had: (i) showing that the fit of the real-business-cycle theory with the facts was improved, in comparison with the fit of its prototype version, when neglected features of the economic transmission mechanism were taken into account; (ii) showing that it was possible to introduce not only real shocks but also public demand shocks in the analysis, particularly for the study of the effectiveness of monetary policy; (iii) showing that it was possible to substitute monopolistic competition for perfect competition in the artificial economy; (iv) showing that it was even possible to introduce other forms of non-competitive behaviour, or market frictions or price rigidities, and to so deviate from complete market clearing. The survey will lead us to quote six chapters of the book used here as the main reference, after the quotations given earlier of the introductory chapter[65].

1. Many commentators observed that the main deficiencies of the simple model of the real business cycle in the test given by stylized facts concerned *the labour market*, even when stylized facts ignored unemployment: real wages were in fact less volatile than the model predicted, in comparison to either employment or labour productivity; productivity was in fact a leading indicator, which did not agree with the model. Writing in the Cooley book under the title “Business Cycles and Aggregate Labor Market Fluctuations”, F. Kydland states two main objectives: first, “to give examples of the perceived deviations relative to theory, especially those related to labor market fluctuations, and of how researchers have attempted to resolve them”, second, “to present in detail… a model environment that is reasonably rich in its description of the labor market”.

A first line of research concerned intertemporal substitutions in the supply of labour to firms. It was argued that these substitutions, induced by changes in the real wage, did not result from the choice between working and enjoying leisure, but also from the choice between working for pay and working at home. Modelling household production then appeared as both valuable in

[65] The book contains twelve contributions. One concerns policy analysis. Two others deal with computing, which would take us too far afield here. The two last contributions concern parts of macroeconomics which our book does not pretend to systematically cover, as was indicated from the start: international economics and financial economics.

itself and as directly relevant for tests of the real-business-cycle theory. A special chapter of the Cooley–Prescott book, by J. Greenwood, R. Rogerson and R. Wright, is devoted to home production, beyond what is reported in Kydland's chapter. Contribution of this development to the explanation of labour market facts hinges on three parameters: the standard deviation of the home-technology shock, its correlation with the more common shock concerning business-technology, and crucially the (short-run) elasticity of substitution in preferences between market- and home-produced consumption. Proponents of the part played by household production in explaining business cycles accept a surprisingly high level of the latter elasticity.

Starting from the fact that the number of hours worked by low-wage employees is much more volatile along the cycle than the number worked by high-wage ones, Kydland reports results obtained with a model in which two representative individuals exist, with respectively high and low skill, the objective function replacing (90) being a weighted sum of the utilities of the two individuals. The model can explain the fact in question, at least if substantial transfers of income take place from the high-skilled to the low-skilled individual.

Any accurate description of labour market facts related to business fluctuations has to distinguish hours worked on average per worker from employment: as we saw, hours per worker are slightly leading output, employment is lagging. The direct interpretation of the prototype model would say that fluctuations in the labour supply mean fluctuations in hours per worker. In order to explain separately fluctuations in employment and in hours per worker, the model has to capture the fact that two decisions are taken at the microeconomic level: whether to work or not? If to work, for how many hours? Kydland presents a model intended to achieve such a result. Earlier G. Hansen had proposed a model with "indivisible labour" in which hours per worker were exogenously given and there was, at the microeconomic level, a fixed cost for working; decisions of the representative individual were then aggregating microeconomic decisions of people working and people not working[66]. Concluding his chapter, in which other contributions are also discussed, Kydland writes: "In the initial development and use of this framework, some features. . ., especially the volatility of aggregate hours of work and the correlation between hours and productivity, were regarded as important deviations [of data] relative the theory. As theory and measurements have progressed, however, the status of these features as deviations has diminished".

[66] G. Hansen, Indivisible labor and the business cycle, *Journal of Monetary Economics* **16** (1985) 309–327.

2. We now turn attention to the chapter entitled "*Money and the Business Cycle*", written by T. Cooley and G. Hansen. We neglect the model meant to capture confusion in the interpretation of an imperfect signal received by agents about the shocks: this model transposes to technological shocks the type of argument explained here in Section 5.2 of Chapter 8 about the effect of imperfectly observed monetary shocks; Cooley and Hansen conclude that "introducing imperfectly observed technology shocks into a real business cycle model does not significantly change the business cycle properties of the model". The chapter presents two other models in succession, but in our brief examination here we can go directly to the last one, which simply adds an important new complication to the specification of the preceding one; the complication is important because it captures a source of nominal rigidity and because it leads to a significantly better match with the data.

Money is required in the model in question for the purchase of a consumption good. According to the "cash-in-advance" assumption, the representative individual must have available at the beginning of period t a money balance covering the value of this consumption good, to be delivered and consumed at the end of the period. As output of production, this consumption good is perfectly substitutable in consumption to the other good. Technological shocks ε_t appear in the production function, as they do in other models of the RBC movement. The supply of money grows according to a fixed exponential trend, subject, however, to monetary shocks μ_t. Increases in the quantity of money finance current public expenditures, which for simplicity are assumed to be just transfers to the individual. This device implies that monetary shocks are transmitted to the nominal wealth of the representative individual, hence to its future flow of nominal demands for consumption. But real demands will depend on the price of the good (serving for capital and, in two forms, for consumption), a price that is determined in each period by the condition that the market clears in that period.

The market for labour is a little more complex because the nominal wage rate for period t is agreed one period in advance. In period $t-1$ the nominal wage rate for period t is competitively determined on the basis of expectations about the technology and monetary shocks (ε_t, μ_t) which will occur in period t; this wage rate then depends on $(\varepsilon_{t-1}, \mu_{t-1})$. In period t, after ε_t and μ_t are revealed, the firm chooses the labour input to be paid at the previously agreed nominal wage; the individual, taking the demand for labour as given, chooses consumption, investment and the increase in its money balance.

With such a model properties of the equilibrium deviate, in an important respect, from what they were in the prototype and other early real-business-cycle models. Because of the cash-in-advance constraint, we no longer associate to the competitive equilibrium an optimality property; we can no longer speak of implementing an optimal programme; we cannot determine the competitive equilibrium by direct application of procedures serving for the computation of optimal programmes. Nevertheless neither the approach nor the computing procedure is wholly different, as shown in the Cooley–Prescott book.

When the model is calibrated, solved and compared with the data, it appears that monetary shocks have significant effects on the cyclical properties of the solution. Some of these effects go in the required direction for a good fit with the data: in particular, the rate of inflation and nominal interest rates are both positively correlated with output, with correlation coefficients of a similar magnitude to those found in the US stylized facts. But other effects lead to comovements that are inconsistent with what is observed in US data: in particular the correlation coefficients with the rate of money growth are, for most variables, quite different in the model from what they are in the data.

3. J. Rotemberg and M. Woodford wrote the long chapter devoted to the consideration of *imperfectly competitive product markets*. By its pattern and by the aspects it emphasizes the chapter somewhat deviates from others appearing in the same book; we may wonder whether the authors should be ranked as taking part in the RBC movement[67]. But the subject is important and well belongs to the broad conception of the RBC research programme.

When studying how the economy responds to various shocks at business cycle frequencies, we find good reasons for allowing competition to be imperfect. Perfect competition assumes away increasing-returns technologies, whose presence nevertheless can be established and matters for a complete understanding of cyclical variations, particularly variations in measured labour productivity. Imperfect competition changes the predicted effects of technology shocks when they occur, as well as the predicted effects of other shocks. The range of the latter is enriched by the possibility of exogenous shocks affecting markups and prices.

Rotemberg and Woodford deal with monopolistic competition, which was introduced here in Section 2.1 of Chapter 4. An important consequence of monopolistic competition is a discrepancy between the marginal productivity and the real wage rate, this discrepancy being related to the price-elasticity of the demand addressed to the representative firm and being all the larger as the absolute value of the elasticity is smaller (see Equation (67) of Chapter 4). In order to integrate this feature in a complete general equilibrium framework, the model has to represent firms as producing differentiated goods. But that can be conveniently done under specific assumptions about technology and

[67] Indeed, see the reference to another article of Rotemberg and Woodford in Subsection 2.5.2.

under homotheticity of the preference map, assumptions which lead to easy price and quantity aggregation[68]. In order to permit a good calibration to actual economies, the model must allow for the presence of intermediate inputs, a point which motivates a serious examination in the chapter and explains particular features of the chosen specification[69]. These features imply that the equilibrium value of the ratio between the marginal productivity of labour and the real wage rate must be equal to a constant μ directly determined from the parameters of the specification and called "the inefficiency wedge" in the chapter.

The authors study the effects of three kinds of shocks: the traditional labour-saving technological shocks, the demand shocks due to government purchases compensated by an opposite change in the net wealth of individuals (the Ricardian equivalence under rational expectations, see here Chapter 6, Section 2.4), the price shocks or mark-up shocks linked to changes in μ. In order to present their numerical results they mainly exhibit impulse-response functions, of the type discussed here in Section 5.6 of Chapter 8, rather than the predicted stylized facts commonly shown by proponents of the movements[70]. They systematically compare the responses to those given by their model for the limit case of perfect competition and constant returns to scale, i.e. $\mu = 1$.

Rotemberg and Woodford begin with the demand shocks. *Under perfect competition and constant returns to scale* an increase in government demand, leads to a decrease in private net wealth and to an increase in the real interest rate, hence to an increase in the endogenous labour supply and to such decrease in the equilibrium real wage rate that employment increases by the required amount. Equilibrium output also increases, although relatively less, because the marginal productivity of labour is smaller than 1 (the "multiplier" also is definitely smaller than 1). After the period of the initial shock, these effects become progressively less important, because in particular the cyclical component of the process of government consumption is drawn by the authors from the fit on a first-order autoregressive process. These responses are interesting to note because they show how demand shocks can be studied in models accepting market-clearing, perfect competition and rational expectations (they

[68] This specification was first introduced by A. Dixit and J. Stiglitz (Monopolistic competition and optimum product diversity, *American Economic Review* **67** (1977) 297–308) and was used since then in the macroeconomics of imperfect competition. See, for instance, O. Blanchard and N. Kiyotaki, Monopolist competition and the effects of aggregate demand, *American Economic Review* (September 1987).

[69] We shall briefly consider implications of the presence of intermediate goods in Section 2.8.

[70] Impulse-response functions are, however, given as additional information elsewhere in the book: in the chapter by Cooley and Hansen, as well as in the chapter on international business cycles.

would indeed be also obtained with the model discussed in Subsection 2.3.2 if only demand shocks would replace technological shocks and labour supply be endogenous)[71].

Results *under monopolistic competition and increasing returns to scale* are qualitatively changed when $\mu = 1$ is replaced by $\mu > 1$; they are worth considering. Choosing $\mu = 1.4$ as the calibrated value, the authors find that output responses are increased by 25 per cent about; on the contrary the impact on the labour inputs and on real wages are each smaller by about 10 per cent under monopolistic than under perfect competition. On the whole the imperfection of competition has just a small impact as long as we keep market clearing and rational expectations. This is precisely the conclusion we reached at the end of the static analysis of Subsections 2 and 3 of Section 6.2 in Chapter 7.

Effects of technology shocks under monopolistic competition were studied not only by Rotemberg and Woodford but also by A. Hornstein, who used a slightly simpler model and presented his results according to the standard pattern promoted by the RBC movement[72]. Hornstein showed that, with reference to the standard model, introduction of both monopolistic competition and increasing returns to scale had effects which were sensitive to the precise choice of various parameters; in some cases output volatility was unaffected. But Rotemberg and Woodford argued that the response of output ought to be higher under imperfect competition, the difference being however small in their numerical results. The two investigations agree in concluding that fluctuations in labour input are smaller with imperfect competition.

The three authors agree also in insisting on the point that, under monopolistic competition and increasing returns to scale, empirical Solow residuals overstate the importance of technology shocks when they are taken as measuring these shocks[73]. Indeed, shocks in government purchases alone generate changes in Solow residuals. Even when technology shocks alone occur, there is an upward bias: the standard deviation of the true shocks may then be grossly overestimated.

Studying how the economy reacts to exogenous mark-up shocks, Rotemberg and Woodford provide estimates of the responses. Given their calibration, a 1

[71] For introduction of demand shocks (on government consumption) simultaneously with technology shocks in the standard RBC model see L. Christiano and M. Eichenbaum, Current real business cycle theories and aggregate labor-market fluctuations, *American Economic Review* (June 1992). The main message is that their introduction of demand shocks decreases the correlation between labour input and labour productivity; it so brings the model closer to reality.

[72] A. Hornstein, Monopolistic competition, increasing returns to scale, and the importance of productivity shocks, *Journal of Monetary Economics* **31** (1993) 299–316.

[73] The point had been made earlier by R. Hall, for instance, in: The relation between price and marginal cost in US industry, *Journal of Political Economy* **96** (1988) 921–947.

percent exogenous temporary increase in the mark-up indicator μ induces a decrease of some 0.4 per cent in the real wage, hence a decrease of 1.4 per cent in the labour supply, leading to a decrease of 1.1 per cent in output.

4. In their chapter on "*Non-Walrasian Economies*", J.-P. Danthine and J. Donaldson show how the RBC methodology can be applied to economies in which some market transactions are effected at non-clearing prices. By so doing they want to demonstrate that applications of the methodology need not have "a new classical bias". In particular applications may account for the empirical characteristics of unemployment in modern economies: (i) a fraction of the labour force is unemployed at any given time, (ii) this fraction varies over time, and (iii) it assumes different average values in different economies. The authors claim that their chapter shows how the RBC canonical model can be adapted to accommodate a variety of non-Walrasian features.

Their adaptation distinguishes two categories of agents in the households sector. Shareholders own the firms, consume, invest, but do not work. Workers have no capital, except for their uninsurable human capital. For simplicity their labour supply is assumed to be fixed. Shareholders are represented by an infinitely lived maximizing agent. Workers, who consume their current income in each period, are likely to experience substantial income risk. But, so the authors argue, "modern economies have developed substitute mechanisms for smoothing consumption"; it is possible "to demonstrate that this enlarged role of labour institutions and arrangements is not without consequences for the dynamics of the economy". More precisely, the authors present models meant to capture one or the other of two ideas found elsewhere in the literature: segmented labour markets (with "primary" and "secondary" workers, or with "insiders" and "outsiders"), and efficiency wage contracts.

In the presentation here we may limit attention to a simplified version of one of the models discussed. There are two categories of workers. Workers in the primary sector are permanent members of the firms, are always employed and their utility is incorporated with the objective function of shareholders (with however a smaller weight). Workers in the secondary sector are employed or not, and with contracts for just one period. The real wage paid to permanent workers is determined by shareholders in their altruistic maximization. Temporary workers must be paid at least an exogenous minimum real wage; in good times, they are all employed and paid a higher market-clearing real wage.

Danthine and Donaldson show that, with such non-Walrasian models, employment is found to be more volatile than with the common Walrasian models, a feature which improves the match with empirical facts. They close their chapter by pointing out that non Walrasian models are not limited to the recognition of labour market frictions. Non-clearing product markets could be incorporated and, still more promising, rationing on credit markets.

The models discussed by Danthine and Donaldson contain, as we just saw, more agents than do the standard models most often seen as coming from the RBC movement. Other authors also have considered such cases, as is pointed out in a separate chapter of the book: J.-V. Rios-Rull, "*Models with heterogeneous agents*". Rios-Rull shows that, in some theoretical constructs, heterogeneity goes with the existence of uninsurable risks, like in Danthine–Donaldson. But emphasis may then be placed on a different kind of stochastic impulses: individual risks, also called "idiosyncratic risks", rather than on an aggregate risk similarly affecting all firms or all households. Emphasis may also be placed on the fact that, credit markets being incomplete, agents are often faced with liquidity constraints.

Another important source of heterogeneity in models of intertemporal allocation comes from the fact that several generations of agents are simultaneously living and exchanging between them. As we saw in our study of economic growth, substituting an overlapping-generations model to a model with infinitely lived agents can change the dynamic properties of the equilibrium path (compare Sections 4.2 and 5.2 to 5.4 in Chapter 6). Rios-Rull also discusses how overlapping-generations models can enter into the RBC research programme.

2.5. Dynamic stochastic general equilibrium: research and perspectives

We shall now go on with a change of perspective, still looking at the same object, the RBC approach, but no longer from within the movement. Our main query will be whether the approach launched with the so-called theory of real business cycles is appropriate for a good understanding of business fluctuations. Is it appropriate for gauging the importance of technological shocks? Is it appropriate for a good representation of the propagation mechanism? Is it inherently biased against the recognition of a number of facts and for the promotion of a number of ideas? Is not the history of business cycles marked by persistent market malfunctionings which at times occurred and will always look foreign to the preconceptions embedded in the approach? Will the top-down strategy favoured by the movement be cost-effective for research on a quite complex set of phenomena?

With this query in mind we shall consider a few recent publications which, for one reason or another, seem to be less germane than those surveyed in the preceding section to the initial ideas of promoters of the RBC approach. But the reader should not misinterpret what is mainly a convenient device for a thought-provoking shift in perspective. It would indeed be artificial to mark a neat frontier separating research done within the RBC movement from what could now be ranked here under the label of "related research". The book edited by T. Cooley provided an easy reference for delineating what belonged to the movement, but the works to be now considered do not stand at such a larger distance from the real business cycle theory that at least some of them could not conceivably have appeared in the book.

2.5.1. *The Solow residuals*

We begin with an article by C. Evans questioning the exogeneity of the Solow residuals with respect to the business cycle[74]. This is an econometric study applying a test of Granger non-causality (see Chapter 8, Section 5.4) to the US quarterly series of Solow residuals for the period 1957 to 1983. The maintained hypothesis was that, simultaneously with the Solow residual, a number of other variables were ruled by a multidimensional stationary autoregressive process. The hypothesis to be tested was whether there was no significant evidence that these variables would cause, in the Granger sense, the evolution of the Solow residual. Indeed, absence of causality is assumed in all applications where innovations to the Solow residuals are taken as exogenous shocks. Evans considered the following variables: the quantity of money, a short-term interest rate, the consumer price index, real government expenditures and the price of crude oil. For each of these variables taken individually, except for the oil price, the test concluded that there was significant causality from the variable to the residual.

Evans also proceeded to a variance analysis in order to see how much of the 16-quarter-ahead forecast error variance of the residual was attributable to the additional variables. He concluded from his analysis that between one quarter and one half of the variance was "attributable to variations in aggregate demand", implicitly admitting that the rest could be attributable to variations in technology. This last conclusion is disputable. Indeed, there are other causes of variations in aggregate demand that what is captured by the variables considered: variations in exports, in the pressure of taxes, in the confidence of business firms and households, ... Moreover, the results of Rotemberg–Woodford and of Hornstein led us to conclude that the variance of the Solow residuals might overstate the variance of technological shocks, even if such shocks would be the only ones to matter.

[74] C. Evans, Productivity shocks and real business cycles, *Journal of Monetary Economics* **29** (1992) 191–208.

The issue so raised prompts us to consider the article by C. Burnside, M. Eichenbaum and S. Rebelo[75], who write in their summary: "We estimate that the variance of innovations to technology is roughly 50 percent less than that implied by standard real business cycle models; ... existing real business cycle studies substantially overstate the extent to which technology shocks account for the variability of postwar aggregate US output". The article has two attractive features: its model recognizes the presence of another kind of shocks besides technology shocks, namely those affecting government consumption (an imperfect representation, however, of aggregate demand shocks); the model focuses attention on labour hoarding, which was long recognized as being a main determinant of the Solow residual.

But we may wonder whether their model is fully realistic. It is an elaboration of Hansen's model of "indivisible labour". Adjustment costs, usually meant to be the main reason for labour hoarding, do not explicitly appear. Their effect is implicitly captured by the assumption that employment is decided not instantaneously, but rather one quarter before it will be used in production and before the then relevant shocks will be observed. When these shocks will be known, production will be decided simultaneously with work effort. The new variable, work effort, appears in both the utility and the production functions. The representative individual, or rather both the representative worker and the representative firm, optimally decide in each period what effort should be put in work, at the same time as they choose their labour supply and demand. This modelling device is clever if our objective is to fit labour hoarding into the optimal growth framework. But does it lead to a persuasive model and to reliable estimations? We observe in particular that the short-run sensitivity of effort to shocks implies a concordant short-run sensitivity of the market-clearing real wage, but the latter is not found in the data.

2.5.2. *A weak propagation mechanism*

As we saw in the third part of Section 2.3, the good match with stylized facts that is commonly claimed in applications of the RBC approach does not provide a powerful test for theories of business fluctuations. Some recent contributions illustrate this remark, particularly an article by T. Cogley and J. Nason[76], where it is found that many models built by the RBC movement have weak endogenous propagation mechanisms and do not generate interesting business

[75] C. Burnside, M. Eichenbaum and S. Rebelo, Labor hoarding and the business cycle, *Journal of Political Economy* **101** (1993) 245–273.

[76] T. Cogley and J. Nason, Output dynamics in real-business-cycle models, *American Economic Review* (June 1995).

cycle dynamics via their internal structure[77]. They therefore have to rely heavily on exogenous sources of dynamics in order to replicate some stylized facts.

The authors pay attention to two facts: first, GNP *growth* is positively autocorrelated over short horizons and has weak and possibly insignificant negative autocorrelation over longer horizons, second, GNP itself appears to have a strong trend-reverting component. More precisely the authors take as a benchmark some statistics characterizing the quarterly series of US real per capita GNP over the years 1954 to 1988: the correlogram and the smoothed periodogram of the growth rate, two impulse-response functions for respectively a permanent and a transitory component of the per capita GNP series. The decomposition results from application of the technique worked out by O. Blanchard and D. Quah (see Chapter 8, Section 7.5) with just the difference that it is based on the simultaneous observation of per capita labour input, rather than the unemployment rate, together with per capita GNP. For each component are given responses, after successive lags, to the impulse due to its own innovation.

The correlogram has values of about 0.3 and 0.2 for the two first lags; at higher lags the autocorrelations are mostly negative and statistically insignificant. The smoothed periodogram has a broad peak that ranges from 2.33 to 7 years per cycle. The impulse-response function of the permanent component increases at a decreasing speed (what the authors call the "hump-shape"), being equal to approximately twice the impulse after 5 quarters, three times after 12 quarters and finally levelling at about four times. The impulse-response of the transitory component increases during the first three quarters, up to about 1.5 times the impulse, then slowly decreases to zero (still equal to about 0.3 times the impulse after five years). The transitory component is not negligible since its impulses are on average more than four times larger than those of the permanent component.

Such are the facts identified by Cogley and Nason about the dynamics of US real per capita GNP. In order to know whether these facts agree with the dynamics induced by the propagation mechanisms of various RBC models, the authors proceed to a Monte Carlo simulation study of the output series generated by these models when their shocks follow random walks. The results are distressing: models relying entirely on capital accumulation and intertemporal substitution lead to results with hardly any dynamic feature: no autocorrelation, a flat smoothed periodogram, a negligible transitory component and a non-increasing impulse-response function of the permanent component, which means no important trend-reverting component; models with gestation

[77] This point was earlier made, about the standard real-business-cycle model presented by T. Cooley and E. Prescott (1995, op. cit.), in an article to which me shall refer in Section 4.5: M. Watson, Measures of fit for calibrated models, *Journal of Political Economy* (December 1993).

lags or adjustment costs of the capital stock do hardly better; judging in particular from results obtained with the model built by Burnside, Eichenbaum and Rebelo (op. cit.), employment lags or adjustment costs of the labour input do somewhat better, with a built-in short-term autocorrelation and a periodogram decreasing for high frequencies; but their transmission mechanism implies no trend-reversion and understates the response to transitory shocks.

Cogley and Nason then study what kind of dynamic properties of the shock processes contained in the models would be needed for a good replication of the stylized facts they identified. With a sufficient number of degrees of freedom many possibilities are open. But the models may still impose strange conditions. For instance, on a model to which we referred earlier (L. Christiano and M. Eichenbaum, 1992), Cogley and Nason have to introduce, in the autoregressive shock process of government spending, coefficients that seem problematic in comparison with what a direct study of this process suggests. Even so, the replication of stylized facts is imperfect.

The same kind of critique of the usual empirical validation of RBC models is issued by J. Rotemberg and M. Woodford[78], who even challenge the initial intuition that founded the theory of real business cycles. They write: “One of the principal attractions of the RBC model is its parsimony. The model is supposed to explain growth that is simultaneous with the business cycle using only one set of shocks... We show that, although the model is constructed so that it can explain the stochastic trend in output, it is unable to account for the business cycle as we define it”.

The empirical evidence used by the authors is drawn from a three variable vector autoregression on quarterly US data; it consists of the forecastable movements in output, consumption and labour input. The three variables for the VAR are chosen so that their process can be assumed stationary: the growth rate of output, the log of the ratio of consumption to output and the log of the detrended labour input. From the results it is possible to compute interesting statistics, such as the standard deviation of the expectation at time t of the cumulative change in log output from t to $t + \tau$, and this for successive values of τ (the forecastable movements in output at the τ-horizons). It is observed in particular that forecastable movements in the three variables are strongly positively correlated.

Rotemberg and Woodford show that the cyclical features so characterized do not agree with those of the forecastable movements generated by the standard RBC model with technology shocks following a random walk: the variance

[78] J. Rotemberg and M. Woodford, Real business-cycle models and the forecastable movements of output, hours and consumption, *American Economic Review* (March 1996).

of the expected movement in output over the next four quarters is about 1 per cent of what is found from the data, consumption and labour input are typically moving in opposite directions according to the model, the correlations between the expected movements of the three variables are weak in the model. Alternative specifications do not remove the problems. In particular additional disturbances would have to exhibit persistence over many quarters if they were to explain the facts, their cyclical properties then dominating the properties induced by the assumed propagation mechanism.

2.5.3. *Flows on the labour market*

A number of recent contributions pay attention to features of the labour market. After what was reported earlier about the chapters by Kydland, Danthine-Donaldson and Rios-Rull in the book edited by T. Cooley, we shall not systematically survey these contributions. But we shall briefly look at an article by M. Merz[79], which is exceptional because it deals with search and flows on the labour market and because it dares to consider the unemployment rate as a significant macroeconomic variable.

In order to replace the frictionless Walrasian labour market by one in which trade frictions are present, the author integrates, into an RBC model with just technological shocks, a representation of labour-market flows and search, then accepting the approach discussed in Section 8.4 of Chapter 8. Denoting job vacancies as V_t, matches on the market as M_t, employment, or equivalently labour input, as L_t and unemployment as $1 - L_t$ thanks to a normalization making the labour force equal to 1, the author introduces the matching function:

$$M_t = V_t^{1-\lambda}\big[S_t(1 - L_t)\big]^{\lambda}, \quad 0 < \lambda < 1, \tag{106}$$

where S_t is a measure of the intensity of search devoted by an unemployed worker to find a job. Searching jobs and posting job offers is costly, in conditions which remain unchanged when technical progress occurs according to trend in the production sector: in other words, for given L_t, S_t and V_t, a fixed share of trend production is used in these activities; equilibrium in the market for goods so implies:

$$C_t + I_t + e^{\nu t}\big[C(S_t)(1 - L_t) + aV_t\big] = y_t, \tag{107}$$

[79] M. Merz, Search in the labor market and the real business cycle, *Journal of Monetary Economics* (November 1995).

where a is a positive parameter and $C(S_t)$ the search cost function. Finally a fixed proportion ψ of employed workers is assumed to leave employment during each period, so that:

$$L_{t+1} = (1 - \psi)L_t + M_t. \tag{108}$$

So, the same kind of link exists between the series of M_t and that of L_t, as is traditionally written between the series of investment I_t and capital K_t. Hence, labour input of period t results from decisions taken before knowledge of the technological shock of the period. Except for the above additions and changes, the standard RBC model is maintained.

The planning problem of optimal stochastic growth is then solved with as decision variables of period t (after the technological shift is observed): S_t, V_t, M_t, y_t, I_t, C_t, L_{t+1} and K_t, starting from the initial values L_1 and K_0 (we stick here to the timing introduced in the expression of the production function (85)). In the solution, a positive technological shock leads to high values of V_t, hence M_t, hence L_{t+1}. Moreover the cost of posting vacancies dampens the adaptation since the shock process is assumed stationary and autoregressive. So, both lag and damping are introduced in the evolution of employment.

The model replicates noteworthy features which actually characterize the dynamic behaviour of the labour market but are not replicated by many RCB models: labour productivity is leading employment and is more volatile than real wages; the labour share of income is countercyclical. However, given the focus on labour market flows and stocks, it is disturbing to see that the low level of correlation between the unemployment rate and the vacancy rate does not replicate the short-run counter-movements known as "the Beveridge curve". Would introducing aggregate demand shocks and price rigidity take care of the difficulty?

An improvement in the fit to business cycle facts also follows from alternative changes in the representation of the labour market, as we saw with the chapter written by Danthine and Donaldson in the book edited by T. Cooley. A noteworthy research in the same spirit is due to M. Boldrine and M. Horvath[80]. Some workers are employed in the framework of contingent long-term contracts, which specify up to some horizon the hours of work and the wage, depending on the realization of some random events. Such contracts act as a kind of insurance provided by the employers to the more risk-averse employees. But the exact terms of the contracts depend on the state of the economy

[80] M. Boldrine and M. Horvath, Labor contracts and business cycles, *Journal of Political Economy* (October 1995).

when they are agreed upon. In a recession workers accept terms corresponding to a low reservation utility. When, following an upturn, favourable shocks are realized, entrepreneurs reap most of the benefit because the standing contracts specify that, given the situation, long hours will be worked and the wage, although relatively good, will be smaller than the marginal productivity of labour. It follows that wages paid will be less volatile than productivity and the labour share of income will be low at times of booms. The force of the phenomenon will, of course, depend on how quickly do the terms of long-term contracts adapt to changes in business conditions.

2.5.4. *Price and wage rigidities*

From the static analysis of Chapter 7 we know that price and wage rigidities, due, for instance, to adjustment costs, would also play a significant role. This is well shown by J.-O. Hairault and F. Portier[81] who study an RBC model with monopolistic competition, price adjustment costs, money in the utility function, monetary shocks, and technological shocks which are distinguished from Solow residuals. In order to identify those shocks, the Solow residuals have to be "purged" in conformity with the model. Two calibrations, respectively on US and French data, lead to two distinct sets of simulations. Actually, differences between the two countries appear qualitatively small. More important are the improvements in the match to observed facts with respect to other models. In particular recognition of monetary shocks seem to be necessary if the simulations have to reproduce the standard deviation of output, the money-output correlation, the correlation and relative standard deviations of labour productivity and labour input, and the correlation of mark-ups with output.

The study of the nature and impact of price and wage rigidities within the RBC approach is found relevant and is actively pursued by a number of other research workers. They are motivated by the purpose of clarifying the mechanism by which monetary policy affects real economic activity, a concern that already underlain much of the work reported here in Chapters 7 and 8. In particular T. Cooley and G. Hansen[82] complement their chapter in the book we discussed in Section 2.4. They again study three RBC models that differ according to the mechanism through which monetary growth shocks operate. But the specifications or calibrations are somewhat revised, the sample of quarterly US data running from 1954 to 1994. The main change concerns the model with

[81] J.-O. Hairault and F. Portier, Money, new-Keynesian macroeconomics and the business cycle, *European Economic Review* (December 1993).

[82] T. Cooley and G. Hansen, The role of monetary shocks in equilibrium business cycle theory: three examples, *European Economic Review* **42** (1998) 605–617.

labour contracts extending to more than the current period, precisely the one we stressed in Section 2.4: such contracts now apply only to 42 per cent of employees but fix the wage for four quarters; the wage of other employees is determined on the competitive spot market. The authors find that their two economies with either labour contracting or misperception of monetary shocks have similar cyclical properties and best fit the data. However, these properties and those exhibited by the US economy differ, particularly because the models capture neither the phase shift found in the correlation of money growth with real variables nor the negative correlation between money growth and nominal interest rates. The authors conclude that more work is needed on the monetary transmission mechanism.

O. Jeanne[83] argues in favour of recognizing that, besides some nominal rigidity, real-wage rigidity plays an important role. He first notes that, with a plausible degree of price stickiness, current RBC models lead to the conclusion that money shocks do not produce the level of persistence in economic fluctuations that is observed in the data (Subsection 2.5.2 above). He then claims to show that the conclusion is reversed if there is also moderate real-wage rigidity. Presenting in detail the general equilibrium model here would be too long (perhaps, however, this model will initiate a new and promising strand of dynamic-stochastic-general-equilibrium modelling). Let us just note that the specification of the effect of nominal rigidities brings in back the type of phenomenon that was central in our Chapter 8, starting with Section 2.1. The author is eventually led to find a second-order autoregression for the deviations of output from trend, with a degree of persistence and cyclicality that seems to give a good empirical fit. He concludes with speculations about what might result if real-wage rigidity was better embedded in all the behaviours assumed in his model.

2.5.5. *Multisectoral cycles*

Transmission of technological shocks to all characteristics of the moving temporary equilibrium begins with their effects through the interindustrial input–output relations within the productive system. This initial phase takes time and might conceivably generate cyclical features. It was neglected thus far here. However, one of the two 1983 articles founding real business cycle theory, the one authored by John Long and Charles Plosser, was multisectoral and mainly devoted to the dynamics of the transmission of technological shocks through the interindustrial structure. The introduction of the article lists two

[83] O. Jeanne, Generating real persistent effects of monetary shocks: how much nominal rigidity do we really need?, *European Economic Review* (June 1998).

broad regularities exhibited by business cycles, the second one being: "Most important, measures of various economic activities (e.g., outputs in different sectors) move together". Moreover, we saw at the end of Section 2.3 that the study of time series disaggregated by industries could be interesting in order to identify the joint statistical properties of erratic movements in productivity in various industries: the study might in particular show whether and how such properties fit with the idea that the productivity movements in question would reflect technological shocks. A brief look at the RBC literature on multisectoral cycles is therefore justified here.

1. A convenient entry to the literature is the system of stochastic linear equations generated by the multisectoral log-linear economy, which recognizes interindustrial input–output relations, generalizes the economy discussed in Subsection 2.3.2 and played a major part in the Long–Plosser 1983 article. Without going into an analytical treatment generalizing the one spelled out in Section 2.3, let us simply review the main elements of the model and the system of equations to which it leads.

In Section 2.3 the main elements were the production function (85), the objective function (89) and their particular specification given by (91). The main variables were output y_t, the random technological level u_t, the inputs L_t and K_t, finally consumption C_t. Introducing the interindustrial structure discussed in Chapter 3, we have to identify n industries ($i = 1, 2, \dots, n$), industry i producing in period t the output y_{it} of good i from the inputs $L_{i,t-1}$ of labour and $x_{ij,t-1}$ of the various goods j ($j = 1, 2, \dots, n$), assumed here to be drawn from the outputs $y_{j,t-1}$ of the previous period; similarly u_{it} is the random technological level experienced by industry i and C_{it} is consumption of good i. Generalizing (91), we assume now that, in the objective function, the utility U is a linear function of the n variables $\log C_{it}$ and that, according to the production function of industry i, returns to scale are constant and $\log(y_{it}/u_{it}L_{i,t-1})$ is a linear function of the n variables $\log(x_{ij,t-1}/u_{it}L_{i,t-1})$ for $j = 1, 2, \dots, n$. We may say that all production functions are Cobb–Douglas and that all elasticities of substitution are equal to 1. We note in passing the universal one-period lag between inputs and their immediate output. (In applications the assumption and the length of the unit period ought to be discussed.)

An analytical argument may then be given in order to characterize the optimal growth, along lines similar to those used in Section 2.3. We then reached the conclusion that the relevant ratio K_t/C_t depended only, in optimal growth, on exogenous parameters; we were so led to a simple linear autoregressive equation which characterized the growth dynamics, a kind of "reduced equation" ruling this dynamics, with in the background a contemporaneous adaptation of real factor costs. Similarly with the multisectoral log-linear economy, a number of relevant ratios such as x_{ijt}/L_{it} or C_{it}/y_{it} are proved to take, in optimal growth, values than can be directly computed from the parameters; it can then be proved that a "reduced" system of n equations rules the dynamics of optimal growth, with in the background a contemporaneous adaptation of all relative prices. Using $\log y_t$ as the notation for the n-vector with components $\log y_{it}$ we may write the system in question as:

$$\log y_t = A \log y_{t-1} + k + \log u_t, \tag{109}$$

where the matrix A and the vector k depend only the parameters of the model (the elements of A are equilibrium cost shares in the production of each output; they depend on the parameters of the production functions; the elements of k depend on parameters of the objective function). The system is remarkably simple in showing how the stochastic dynamics of the technological levels u_{it} rules the stochastic dynamics of outputs.

This simplicity is bought at a high cost, namely very special assumptions about production lags, about elasticities of substitution and about homotheticity, in both production and consumption. Whether this cost can be accepted in an application of the model has to be decided in each case. But when simplicity brings with it transparency as it does now, the cost may be worth accepting in a textbook. In this spirit let us go on.

In order to fully characterize the dynamics of the system we need to be precise about the stochastic process of the u_{it}. In Section 2.3 we posed the second equation (99), according to which $\log u_t$ was the sum of a linear deterministic trend and of the random variable ε_t, which had a zero mean value. A similar assumption would recognize specific deterministic trends for the technological levels of the various industries and consider the random deviations ε_{it} of the $\log u_{it}$ from their trends. Dealing with such a case would pose no serious problem, since the argument would transpose the one given in Section 2.3. But, again for simplicity, we are going to follow for the moment Long and Plosser in neglecting the deterministic trends and directly considering the n-dimensional stochastic process $\{\varepsilon_t\}$ whose components $\varepsilon_{it} = \log u_{it}$ are assumed to have zero mean. Alternatively we may say that the log-linear economy leading to Equation (109) is assumed to hold for relations between detrended series.

Long and Plosser consider successively two specific hypotheses about the stochastic process $\{\varepsilon_t\}$. They first assume that the process is made of independent and identically distributed random variables having a unit variance (in this case ε_{it} is uncorrelated with ε_{jt} when $j \neq i$, and also uncorrelated with $\varepsilon_{i,t-\tau}$ when $\tau > 0$). They then point out that the hypothesis deliberately places a heavy burden on the log-linear multisectoral structure in terms of its ability to explain business cycles. After calibration of the parameters to US data they compute and exhibit impulse-response functions of y_{jt} to such shocks as ε_{it}. The results are interesting, but not so favourable to the claim of explaining business cycles: although comovement among different series appears, the response profiles of aggregate output are after all shocks uniformly decreasing.

Long and Plosser also consider the case in which the n variables ε_{it} would follow independent random walks. Using otherwise the same calibration of parameters they present the result of a simulation for hundred successive periods. There is a fair degree of comovement between six of the seven industry series.

2. In what will follow here we shall pose a more general hypothesis encompassing the two considered by Long and Plosser. We shall do it with two purposes in mind. First, we would like to learn about the empirical features of the simultaneous movements of industry output series, looking perhaps at residuals of fits to a system like (109). Second, after a few other economists, we suspect that such an empirical evaluation might be revealing about the respective importances of technological shocks and aggregate demand shocks in business cycles.

Indeed, a system like (109) holds for the log-linear economy under slightly different sets of assumptions. It holds in particular when an autonomous demand is added, for instance, with public expenditures financed for a levy on consumers[84]. If there are shocks in the volume and composition of autonomous demand, together or not with shocks in the technology, a new interpretation must be given of the term $\log u_t$ in Equation (109). It has to capture, also or uniquely, such shocks in autonomous demand.

A number of economists put forward the heuristic idea that in (109) the irregular short-run movements of the $\log u_{it}$ ought to be "industry specific" if they were due to technological changes, being occurring mainly at a time in an industry at the next time in another. In contrast the short-run irregular movements of the $\log u_{it}$ ought to exhibit always about the same commodity composition if they were due to changes in autonomous demand: they should then move up and down simultaneously for the outputs of all industries and be particularly important for those goods used by the public sector. The heuristic idea was indeed underlying the interpretation given at the end of Section 2.3 to the results reached by D. Costello.

Clearly, we cannot claim on *a priori* ground that all forms of technical progress are industry specific: introduction of computers for accounting was

[84] Such an addition is found in some of the contributions reported here in Section 2.4, for instance in the chapter by Cooley and Hansen briefly described in Subsection 2.4.2. Its exact insertion in the stochastic log-linear economy is probably available somewhere in the literature and can also be worked out by the reader.

not. Neither can we claim that accelerations or decelerations of public expenditures always have the same commodity composition. However, if industry specificity was found to dominate in short-run irregular movements of $\log u_t$, this would be a result supporting a key assumption of real business cycle theory. If in contrast little industry specificity was found, the result would rather support the assumption that a key factor in business cycles was changes in autonomous components of aggregate demand. This is why tests of industry specificity of shocks would be significant for the orientation of business cycle theories.

The empirical investigation has, however, to cope with a difficulty, among others. Nobody would deny that long-run trends ought to be industry specific. It is well known that technical progress was particularly fast in those sectors of manufacturing industries producing equipments and particularly slow in many services. Long-run industry-specificity is quite consistent with "the neo-classical synthesis", which limited the role of aggregate demand analysis to the study of the short run. An additional reason for keeping in mind long-run industry-specificity in applications accepting system (109) comes from a clear weakness of this system: being derived from a model with a homothetic objective function, it overlooks the shifts in the composition of consumption as aggregate output and income grow, a shift in favour of "luxuries" as against "necessities".

A way to take the difficulty into account is to assume that $\log u_{it}$ contains an industry-specific deterministic linear trend, as the direct transposition of (99) would lead us to do. We know however, since the influential article of Nelson and Plosser (op. cit.), that this might not completely cope with the difficulty because the long-run trends of some macroeconomic time-series seem to be better represented by stochastic processes of the random-walk type than by linear functions of time, or still because those trends are best represented by a combination of the two. In other words, we have to keep the difficulty in mind even if $\log u_t$ in (109) is replaced by ε_t *plus* a linear trend.

This means that the proper stochastic hypothesis to be made on the multidimensional process $\{\varepsilon_t\}$ has to be wide enough to cover various kinds of interdependence between the dynamic properties of its components. Within the class of multidimensional ARIMA processes we may, however, limit attention to those for which a representation of the following type holds:

$$\varepsilon_t = \sum_{\tau=1}^{h} \Phi_\tau \varepsilon_{t-\tau} + \eta_t, \tag{110}$$

where the Φ_τ are square matrices of coefficients and $\{\eta_t\}$ is a white noise process with covariance matrix Ω.

We may note that the two cases studied by Long and Plosser belong to the general class defined by (110). Both assume Ω to be the identity matrix I: in the first case all matrices Φ_τ are null; in the second case $h = 1$ and $\Phi_1 = I$. But Equation (110) is of course much more general. We also note that, if all matrices Φ_τ were null, characterization of what we mean by industry specificity would only involve the matrix Ω and would be fairly natural. A diagonal matrix Ω of rank 1 would correspond to a clear lack of specificity (knowing a single non-zero component ε_{it} would suffice for knowing "almost surely" what other components happen to be). But, since we want to characterize industry specificity by reference to the short-run (cyclical) dynamics of $\{\varepsilon_t\}$, the first part of the right-hand side in (110) make a relevant definition less simple.

In particular, the process ruled by (110) may have one or several unit roots, which would signal the presence of one or several stochastic trends playing a significant role in the long-run dynamics. Each such trend may be specific to just one sector or may be common to several, even perhaps to all, sectors. But two trends may also be cointegrated, in which case their long-run behaviours would reflect that of the same basic stochastic trend. We thus see that characterizing the "long-run industry-specificity" of shocks is not as simple in general as it is with the second assumption made by Long and Plosser. We may say that, if that assumption holds, full industry specificity of stochastic trends also holds. But in general the degree of industry specificity of stochastic trends ought to be given by the ratio to n of the number of unit roots of the process, a number that may of course be smaller than n, but that may also be smaller than the number of $\log u_{it}$ exhibiting stochastic trends.

The short-run or cyclical dynamics of a process ruled by (110) may similarly exhibit a higher or lower degree of industry specificity. Heuristically we understand that this depends on the set of matrices Φ_τ and Ω, which indeed characterize the full dynamics of the process. The short-run dynamics of the various components may be due to "common features", all characterized by a smaller number of unidimensional basic stochastic processes. An extreme case is the one in which Ω has rank 1 and all Φ_τ are null matrices; but other more interesting cases of lack of short-run industry-specificity can be built. We shall not be more precise in this book in defining such common features, for which the reader may refer to the study to be now examined.

3. After these preliminaries we indeed turn attention to an econometric investigation due to R. Engle and J. Issler[85]. It does not apply exactly what would

[85] Estimating sectoral cycles using cointegration and common features, NBER working paper

follow from a combination of (109) and (110); in particular it does not seem to depend on the validity of the log-linear economy for the modellization of actual input–output relations. But it is motivated by the concern for testing industry specificity in business cycles. The data base is made of annual series of sectoral real GNP in the US, sectoral GNP being divided by the US population, for the years 1947 to 1989; the private productive economy is divided into eight sectors[86]. (Since the GNP of an industry is its value added, not its gross output, input–output relations need not be portrayed.) Let x_{it} be the logarithm of the real value added of industry i in year t *minus* the logarithm of the US population.

The starting point of the analysis is a fit of the series of vectors x_t to a vector autoregression (VAR) to which are added eight linear deterministic trends, one for each component. The subsequent analysis concerns the autoregressive part of the fit, i.e. the process of the vectors x_t purged from its linear trends, $a_i + b_i t$, say.

Engle and Issler find that this fitted process has six characteristic roots which do not significantly differ from 1, but that all eight x_{it} have stochastic trends: it means that two dimensions of cointegration are present; the stochastic trends cannot be taken as fully industry specific although they have a high degree of specificity. The authors estimate the two cointegrating vectors, α_1 and α_2 say. They then graph the evolutions of the two independent stationary basic components $\alpha_1'(x_t - a - bt)$ and $\alpha_2'(x_t - a - bt)$, to be added to six independent basic unidimensional integrated processes in order to form a basis of the multidimensional processes having the same long-run properties as $\{x_t - a - bt\}$. Whereas $\alpha_1'(x_t - a - bt)$ is essentially flat, $x_2'(x_t - a - bt)$ displays oscillations superimposed on two large waves, going the first one from 1947 to 1973, the second one to 1989. Moreover, this second stationary component is found to decrease in every single postwar recession identified by the National Bureau of Economic Research (there were seven such recessions in the 42 years period).

Engle and Issler explain how they derive from their VAR analysis the “common features” of shorter-run movements. They first report that the cofeature rank is found to be six-dimensional: cyclical features of the eight industries so derive from just two independent serially correlated cyclical unidimensional processes. This result already means almost complete lack of industry specificity in cyclical movements, a conclusion that is reinforced by the following closer look.

No 4529 (November 1993).

[86] An econometric investigation of German data is also available in B. Lucke, Productivity shocks in a sectoral real business cycle model for West Germany, *European Economic Review* (February 1998).

It so happens that, in this case, the cointegrating and cofeature ranks add up to exactly the number (8) of variables of the multidimensional process analysed. In such a case the methodology used by Engle and Issler implies that, in the trend-cycle decomposition, the cyclical component of any variable (here $x_{it} - a_i - b_i t$) is a linear combination of the stationary basic components found in the cointegration analysis. Since there are two such components and the first one exhibits very little movement, profiles of the eight industry cyclical components have a great similarity with that of $\alpha_2'(x_t - a - bt)$. Their main differences concern amplitudes: construction has the highest; manufacturing has an amplitude about half smaller, and similarly for mining; the amplitude is one fourth that of construction for the finance sector and one seventh for the trade sector. The three remaining sectors actually exhibit contracyclical movements with small amplitude: one fifth that of construction in agriculture or transportation, a little more than one fourth in services[87].

Roughly speaking, we may say that the picture emerges of idiosyncratic trends in various industries but of a common cyclical movement, very much in agreement with the neoclassical synthesis.

2.5.6. *Reflections*

Let us pause and reflect on the point where we now stand about the short-run macroeconomic dynamics, after Chapter 7 and 8, and after what was already surveyed in this chapter. In the two preceding chapters existence of market tensions and slacks was taken as an essential part of knowledge about this dynamics; changes in tensions or slacks were analysed as inherent to the same phenomenon as fluctuations in unemployment, inflation, productivity and profitability. Fluctuations might originate from monetary shocks, fiscal shocks, shocks in foreign demand, shocks in consumers confidence or entrepreneurs animal spirits, shocks in some sensitive prices or in wage bargaining, even truly supply shocks, such as those induced by sudden shifts in the technology of production. Starting from microeconomic foundations and from aggregation procedures studied in Chapters 1 to 4, econometric investigations were seen as

[87] We must note that the trend-cycle decomposition here made is likely to look a little baffling to economists familiar with the older practice, in which trends were considered as quite smooth deterministic components. In contrast, stochastic trends have somewhat erratic realizations, which in the present study appear, almost all, more volatile than the corresponding series $x_{it} - a_i - b_i t$ themselves (the stochastic trends are then negatively correlated with the associated cyclical components). The authors recognize that their results depend on the particular rule they used for separately identifying the two components. They also give results with two other alternative identifying rules, which make the innovations of the cyclical components orthogonal to that of the associated stochastic trends.

bringing important additional evidence about various behaviours, in various technological, institutional and policy environments. Econometric investigations were also seen as providing the required evidence about the interplay between market tensions or slacks and changes in prices and quantities, as well as about the origins of the shocks and about the policy reactions that they might eventually induce.

The first part of this chapter did not mean a fundamental reorientation of such an approach when it stressed the importance of the investment accelerator, and of capacity limits to expansion, so explaining the likelihood of damped periodicities in business fluctuations. The first section of this second part well fitted within the same broad approach; it simply insisted on the concept of stochastic shock processes, and on the distinction between serial properties of the impulse processes and dynamic properties of the propagation transforming these processes into fluctuations of all variables.

On the contrary, up to this point Sections 2.3, 2.4 and 2.5 may seem to mean that we ought to leave aside all knowledge derived from the approaches studied earlier in the book. The initiators of the RBC movement may indeed have been read as arguing that we could and should start afresh with new and allegedly better founded macroeconomic standards. They may have been read as arguing that macroeconomists could confidently rely on models that would directly and rigorously specify, within an integral and coherent system, the whole economy and its evolution under alternative exogenous hypotheses. The specification would be transposed from the microeconomic theory of intertemporal allocation of resources under uncertainty in an economy made of perfectly rational agents placed in a perfect market system. The transposition would accept simplifying convenient hypotheses, thanks to which the theoretical microeconomic model could be replaced by a computable aggregate stochastic growth model. After calibration and simulations, the results would prove that the aggregate model so posed gave a good approximation to business cycle facts.

A number of economists, particularly among experienced macroeconomists, may have judged the project foolish: such a naive ambition! Such a gamble on its empirical success! Such a daring jump from the theory about the functioning of an ideal economy, a theory whose limitations were already well perceived for its own claimed domain of relevance, to a model to be used in another domain!

Have these reluctant economists been proved wrong? Although limitations of results reached by the new approach now appear, did not macroeconomists learn from the experience? Might not the framework built for the project be now seen as bringing new unexpected relevant evidence? Might not the approach appear promising and likely to be complementary with other approaches? Might not it even eventually succeed better than competing approaches, after a period of experimentation, repair and consolidation?

At present various macroeconomists would give different answers to these questions. Readers of this book will have to form their own judgement about them. In order to assist their reflections, we are now going to refer to the argument developed in a thought-provoking article by J.-P. Danthine[88]. I shall take the liberty to insert in parentheses brief comments or queries of my own as we go along.

2.5.7. *Discussion of a programme*

Danthine starts with the following assertions: "The economics profession is in a state of schizophrenia where most practitioners continue to resort to the neoclassical synthesis for forecasting and policy analysis, while researchers have almost totally discarded it as a framework for conducting research ("almost totally", is it really so? Who is meant to qualify as a researcher? E.M.). Caught in between, teachers often, but not always, present IS-LM to students who are starting economics, while concentrating on its failures at a later stage of the curriculum". (The IS-LM representation clearly exhibits, in aggregate demand analysis, the interplay between the money market and the market for goods and services. As a pedagogical device it stands or fails depending on whether aggregate demand analysis will or not serve students later in their life when they will have to deal with short-run macroeconomic issues. E.M.)

In order to solve the dilemma identified in the two sentences quoted above, Danthine presents his objective as follows: "We want macroeconomics to be a branch of general equilibrium theory, providing not merely a theory of recession or of underemployment equilibria, but principally a theory of cyclical fluctuations, of booms followed by busts, a theory that is inherently dynamic". (Is "general equilibrium theory" a perfectly defined object? If so, which? Is it then ready to absorb the whole of relevant macroeconomics as one of its branches? E.M.) Then Danthine goes on as follows.

"There is a distinct paucity of dynamic general equilibrium models to be used as appropriate platforms for macroeconomics. One can think of two alternative strategies to remedy this deficiency. Following good business tradition, let us label these alternative approaches, bottom up, and top down".

"One line of attack, the *bottom up* approach, consists of starting with the pieces, that is the details of firms' and consumers' problems, and the nooks and crannies of individual markets, focussing naturally on the labour market,

[88] J.-P. Danthine, In search of a successor to IS-LM, *Oxford Review of Economic Policy* **13** (1997) 135–144.

the goods market, and the credit and financial markets... One can then try progressively to assemble the pieces with the objective of achieving, at some later stage, a legitimate dynamic general equilibrium successor to the neoclassical synthesis". (Will this adjective "legitimate" be read by all in the same way? E.M.)

"The other approach, which I label *top down*, consists of starting with full-blown dynamic general equilibrium models (i.e. models with a production sector) capable of providing a complete theory of value(s). Initially, the models are taken straight out of microeconomics and boast few typical macro features (in particular, the first models are moneyless) but this because there [is] such a dearth of dynamic general equilibrium models with macro ambitions. Progressively, one can think of adding specific macro features hoping to connect, at a later stage, with the key observations made by those economists who have invested in the bottom-up approach".

"There is no compelling argument why one line of attack should be viewed as dominating the other. In particular, there is no reason to start with one and wait till the corresponding research programme is complete before embarking on the other". (Indeed, in macroeconomics we cannot long neglect "specific macro features". E.M.) "The RBC programme obviously fits into the *top-down* mode. A lot of what is included in the New Keynesian literature... fits into the bottom-up approach..., but [it] rarely [goes] to the dynamic general equilibrium level which is a prerequisite for a full account of economic fluctuations". (In the last sentence much depends on what is meant by "rarely" and by "dynamic general equilibrium level"; would Hicks' theory of the trade cycle have qualified? E.M.)

Danthine then surveys the successes and shortfalls of the RBC research programme, as he evaluates them. The material often corresponds to what was presented in this section and assessments about it do not differ much from what was written here. At the end he entertains the prospect of "a true Keynesian dynamic general equilibrium model" in which sticky prices and quantity spill-overs would appear.

Two last considerations are stimulated by Danthine's article. He writes in particular: "At the heart of the opposition to the current RBC models, one finds *priors* on the issue of the inherent stability or instability of decentralized economies. The demonstration of such an instability, potentially founded on coordination failures or multiple equilibria,... would reinforce the neo-Keynesian view that business cycle fluctuations are the manifestation of a market failure on a grand scale". We shall indeed turn here to the issue in the next subsection.

In his conclusion Danthine mentions the simplicity of IS-LM and writes: “The key issue [for a successor to IS-LM] may well turn out to be whether the RBC research programme... will be able to deliver a sufficiently simple, transparent and theoretically satisfactory model”. He then frightens an old teacher, who was also a practitioner and who believes in the virtue of transparency, when adding: “It is not inconceivable that the standard tool for macroeconomic analysis, for practitioners and undergraduate students alike, will, in the future, be a palm-top computer version of a sophisticated, friction-prone, dynamic general equilibrium model”. Indeed frightening, since we may fear the confusions which could result, in the minds of students as well as in the minds of users of macroeconomic results so obtained.

A side issue concerns the *top–down* research strategy in the scenario in which the profession would invest more and more efforts into the RBC research programme. Would it be cost effective? Could not less heavy analytical procedures lead to practically equivalent results at lower costs? And we are not speaking for the moment about econometric weaknesses, and others, in applications of the RBC methodology, since the difficulty will be discussed in Part 4 of the chapter.

2.5.8. *Intellectual resistance*

Our reflections need not be bound to the pattern used in Danthine’s article, however interesting it may be. The literature of the last twenty years indeed contains many comments and discussions which are relevant for our present concern. Here we shall just mention three points, which ought not to be ignored after what was seen above. We shall draw the first two of these points from the book of F. Hahn and R. Solow[89]. Our points will concern, first, what is “legitimate” in macroeconomic theory, second, the hypothesis of rational expectations, third, the theory of value in business cycles.

Hahn and Solow were motivated by the intention “to create some sort of respectable theoretical resistance to the main trend of macroeconomy theory in the early 1980s” (p. 1). This trend was identified by the authors as coming from an intellectual commitment to the belief that the only appropriate micro model on which to build macroeconomics was based exclusively on intertemporal utility maximization subject only to budget and technological constraints. It meant that there was “no legal way to talk about pathologies” such as prolonged depression, mass unemployment, persistent inflation. But “those unmentionables needed mentioning”. “In the modern spirit resistance

[89] F. Hahn and R. Solow, *A Critical Essay on Modern Macroeconomic Theory* (MIT Press, Cambridge, MA, 1995).

had to begin with alternative micro foundations" (last three quotations from p. 3). A series of models then showed how macroeconomic theory could be based on alternative and more appropriate foundations[90].

We may indeed wonder how to explain the force that was recognized by many to a position so often found in scientific writings where it is boldly and confidently given as a definitive justification. According to this position there would be no other "legitimate" way to proceed to dynamic macroeconomic analysis except within the confinement of dynamic optimization models of a particular sort. The common wisdom would teach that any model built otherwise would be bound to contain unreliable "reduced equations". We shall see more precisely in Section 3.5 of this chapter why sources of error may damage the conclusions drawn from structural equation models and we shall discuss the conditions under which such errors might be important. But who could seriously claim that no source of error exists in the models commonly worked out under the banner of stochastic dynamic optimization? Where is the proof that these models are immune from specification biases, notwithstanding their high levels of aggregation, the separability properties that they assume, their neglect of many constraints, and so on? It is sound intellectual resistance to pose such questions and to keep looking at all evidence, microeconomic as well as macroeconomic, within a variety of potentially relevant frameworks.

As we realized earlier here, the hypothesis of perfect price and wage flexibility, which was part of the early RBC models, was later revised in some applications of the RBC approach and the prospects given by Danthine about future development of the *top down* approach would reinforce the tendency. But the hypothesis of rational expectations does not seem to have been questioned within the movement.

This is strange, because macroeconomist theorists should feel uncomfortable about the hypothesis as soon as they consider other phenomena than those characterizing alternative scenarios of regular long-run growth. We may grant that exact expectations within each such scenario is a natural and simple hypothesis for such a study of the long run and that it has little chance to be surpassed by a better simple hypothesis. But, as soon as we consider business

[90] The author of this book thought about introducing presentations of these models at appropriate places in his chapters. If he concluded against doing so, it is not only because he was too far advanced into his own work when he read Hahn and Solow. It is also because presentations could not be quick and would have taken place precisely where this book may already appear laborious. In particular the overlapping-generations structure, chosen by Hahn and Solow for most of their analysis, would have meant important additions to the 17 pages of Sections 5.2 to 5.4 of Chapter 6. Similarly, a number of arguments would have had to be connected with those surveyed in Section 7.7 of Chapter 8.

fluctuations neither empirical evidence nor abstract reasoning gives us much confidence in the hypothesis. We looked in Section 1.6 of Chapter 8 at the empirical evidence about the formation of price expectations and saw that it seemed to support neither rational nor adaptive expectations. The empirical analysis of financial markets also raises indirect doubts about the rationality of operators' expectations.

The natural explanation of the realism of exact expectations in the study of regular long-run growth argues that agents must learn by experience what are the permanent trends of the economies in which they are living. But it is not clear at all that the rationalization so given for the long run can be persuasively transposed to the short and medium run. The difficulty comes from the interdependence between, on the one hand, the observations and experiments thanks to which agents learn about the dynamics of the real world and, on the other hand, the determination of short-run evolutions. Even for a good theorist this interdependence is difficult to disentangle. Surveying the research devoted to the question, J.-M. Grandmont summarized the result by stating what he called an "uncertainty principle"[91]. Indeed, a number of examples display enough similarity to suggest that, when agents are somewhat uncertain about the dynamics of the system and are therefore ready to extrapolate a large range of observed regularities, learning is bound to generate instability of self-fulling expectations, unless the influence of expectations on the dynamics is weak. Such results show up in a wide variety of specifications formulating the idea of adaptive learning.

Without going into a study of empirical or theoretical proofs, Hahn and Solow write in their book (p. 140): "The longer the lives of agents, the more incredible the hypothesis that they can predict, or know the stochastic properties of, economic variables over their lifetime. A Ramsey maximizer needs correct expectations out to infinity if there is to be a unique optimum path. It seems to us that no macroeconomic insights can be gained with so extravagant an expectational hypothesis".

Indeed, it is difficult to believe that "bounded rationality" has no relevance for stochastic dynamic macroeconomic analysis. Taking a degenerate form of rationality into account may be still more challenging for the RBC approach than taking price rigidity into account. Another reason for intellectual resistance.

In a sentence quoted in Subsection 7, Danthine defines his top down strategy as involving "full-blown dynamic general equilibrium models... capable

[91] J.-M. Grandmont, Expectations formation and stability of large socio-economic systems, *Econometrica* (July 1998).

of providing a complete theory of value". He certainly means that the general equilibrium models ought to be appropriate to the study of business cycles (indeed, his research with Donaldson is motivated by their concern for appropriateness). Thus, it is fair to wonder about what "complete theory of value" would be appropriate concerning business cycles. This is a challenging question after all previous developments here, from Chapter 7 on.

New classical macroeconomics can be said to have been built on the axiom that the theory of value was a theory of market-clearing prices. But the assumption of full market-clearing is not appropriate for short run macroeconomic analysis or for business cycle analysis. For such purposes we need a theory of value that differs from the one taught in the microeconomic theory of prices and resource allocation, a theory of value coping with important price rigidities and admitting the possible occurrence of autonomous shocks bearing directly on prices (independently of the determinants of supply and demand schedules). Indeed, rigidity is observed in the evolution of such important macroeconomic variables as the real wage rate or the price level. A proper business-cycle theory of value must then incorporate in particular the results of what was learned about the price and wage equations discussed in Chapter 8. In applied macroeconomics some unexplained large movements in the long-term interest rate cannot be ignored and are best viewed as autonomous. These remarks suffice to show the difficulty of the challenge.

Now we may wonder whether a good response to the challenge will be given by the spontaneous development of the research programme in which business cycles are meant to be analysed by dynamic, stochastic, general-equilibrium models. Rather than basing their specifications on the observations concerning the actual rigidity and autonomy of prices, will not research workers adhering to the programme look for new models, each involving one or the other of the many rationalizations which were so far suggested as explaining price rigidity (for instance, a particular form of efficiency wages)[92]? In other words, will efforts be spent on defining new models and deducting their implications, or rather on looking for all relevant facts and on finding how much they agree with each of the now existing models?

In other branches of economics we have examples of prevalence of one or the other of the two approaches. In public economics deducting implications of sensible models dominates in order to study the inefficiencies due to alter-

[92] In Subsection 2.4.4 we briefly presented a model due to Danthine and Donaldson. The model had a segmented labour market, the real wage paid to permanent workers resulted from shareholders' altruism; temporary workers were paid at the legal minimum wage during depressions, at the market-clearing real wage during booms. How close to business cycle facts is such a theory of wages?

native tax-induced sets of distortions, or to study alternative criteria for the choice between public investments. In labour economics interpreting relevant facts is on the contrary the dominant research strategy. The macroeconomics of business cycles clearly stands half-way. Hopefully, this chapter will persuade readers that there are good reasons for keeping the balance between the two approaches.

If so, we should resist the trend toward the inflation of models. We should plea for a critical assessment of the factual validity of the microeconomic underpinnings of each new model, and for a close look at the power of the empirical validation of its macroeconomic results.

2.6. Market clearing and rational expectations: endogenous cycles

We progressively deviated from the initial focus of Section 2.3, which was the study of business fluctuations in an economy with competitive market clearing, agents holding rational expectations. Accepting the approach that gave impulse to the RBC movement, we followed the further developments. But our review of the subject would be incomplete if, before turning to other themes, we did not go back to the initial focus in order to look in the direction of another route, which leads to a completely different vision from the one that inspired the theory of real business cycles. According to this alternative vision, business fluctuations are not simply reproducing the features of the optimal reaction of the economic system to unavoidable exogenous shocks; they are rather endogenously generated within the market system; they are often inefficient and commonly involve volatile expectations; such phenomena appear even with self-fulfilling, hence rational, expectations.

Many contributed, during the past thirty years, to the literature conveying this vision. However, we are not going to enter deeply into the literature in question. In the first place, we already considered it in Chapter 6: after devoting Part 4 to balanced growth, with in particular what we called the "market implementation of optimal growth", we turned in Part 5 to the study of the potentially complex dynamics which may characterize the evolution of a competitive economy; we then met cases of instability, even cases of indeterminacy. Since the overlapping-generations model played a large part, we were then apparently on good ground for the study of growth. But, as we shall see, the difficulties uncovered have a logical nature, are not limited to just a particular class of models and may be responsible for the existence of many sorts of movements, such as, for instance, short-run cycles in an overlapping -generations model with much longer life-cycles.

In the second place, for the time being, the literature on "endogenous cycles" may appear to be weak with respect to the explanation of business cycle facts. Indeed, only few cases exist in which the theoretical models would have been calibrated and confronted to actual data. Whereas the logical possibility of pathologies has been well demonstrated, whereas in some cases the domain within which the possibility would materialize have even been proved to be "large" in some sense, the propositions remain qualitative. Some economists find in this literature reasons to believe in pervasive "coordination failures" undermining the efficiency of the market economy; but other economists take the same outcome of research as negligible; none is actually using it in applications to forecasting or policy analysis.

There are, however, serious reasons for macroeconomists to know the existence of this literature and the nature of its main strands of investigation. It gives rigour to some serious objections raised against theories of the market-clearing rational-expectations school; it so helps whoever wants to know what the domains of relevance of these theories are likely to be. It may succeed in the future in bridging the present gap between its theoretical constructs and the needs of applied macroeconomists. Finally, the literature in question looked seriously into some important aspects of uncertainties in our economies, aspects which should not be ignored by macroeconomists and have not yet been approached in this book.

Here, we shall first briefly survey, for the need of this chapter at this point, part of the research devoted to deterministic non-linear dynamics. This will provide opportunity for adding to references given earlier in the book a few more which may be interesting for readers. We shall, second, look at the interplay between uncertainties and expectations, with in particular the important concept of "sunspot equilibrium" in which self-fulfilling expectations may coexist with market clearing and inefficiency. We shall, third, report some claims recently made about the relevance of endogenous cycles for applied macroeconomics[93].

1. *Deterministic non-linear dynamics* was often used in the theory of business cycles. We examined in particular, in Section 1.3, Hicks's model of the trade cycle published in 1950. Earlier, the interaction of consumption demand and investment demand had already been presented as leading to non-linear systems formalizing the cycle; it was so presented by M. Kalecki, R. Harrod,

[93] For a long survey, covering many more aspects of the subject, see R. Guesnerie and M. Woodford, Endogenous fluctuations, in: J.-J. Laffont (ed.), *Advances in Economic Theory* (Econometric Society Monographs, Sixth World Congress, Cambridge University Press, 1992), Vol. II, pp. 289–412.

N. Kaldor and probably others[94]. The more recent interest in general equilibrium models with explicit modelling of intertemporal optimizing behaviour revived the interest, all the more so as it was contemporary with advances made outside of economics in the mathematics of non-linear dynamic systems. It was then realized that such systems could generate not only cycles but also chaotic trajectories and that the possibility could challenge the vision of good order in macroeconomic evolution.

But a loose reference to the mathematics of non-linear dynamic systems could not suffice for the revelation of significant coordination problems in the market economy. It was also necessary to argue the case within the framework of economic models, which had to be recognized as relevant. The literature that followed is not easily summarized in such a way as to give a good grasp on the problems possibly arising in macroeconomic dynamics. Thus, we have to be content here with an introduction containing a glimpse of the mathematics involved, a glimpse of the theory built by a group of economists and a glimpse of the interface between the two, the purpose being then to convey some intuitive understanding about those mathematical features which are crucial for the validity of the economic argument.

The mathematical question is to characterize the set of trajectories generated by a non-linear law of motion:

$$s_{t+1} = f(s_t) \tag{111}$$

connecting the state s_{t+1} of a system at date $t+1$ to its state s_t a date t. For our purpose the main point is to realize that the dynamics may be quite complex if the sequence $\{s_t\}$ ruled by the law (111) is permitted to experience large jumps, from s_t to s_{t+1}, which would keep occurring, in one direction and another, at ever larger values of t. For instance, in the unidimensional case, we may intuitively understand that this would have a good chance to happen if we knew only that the absolute value of the derivative f' would be larger than 1, time and time again, at points appearing along the evolution. On the contrary, if $|f'|$ remained always smaller than 1 after some time, the trajectory would have to eventually converge to a limit. We shall say no more about the general mathematical treatment of the problem (Guesnerie and Woodford, op. cit., devote eight pages to a presentation of this treatment at the beginning of their survey).

[94] M. Kalecki, A macroeconomic theory of business cycles, *Econometrica* **3** (1935) 327–344; R. Harrod, *The Trade Cycle* (Oxford University Press, 1936); N. Kaldor, A model of the trade cycle, *Economic Journal* **50** (1940) 78–92.

In order to understand how difficulties could occur in intertemporal economic models we shall follow present practice in considering a particular model[95]. This is a simple overlapping-generations model for a one-non-storable-good exchange-economy with stationary fundamentals and an exogenous quantity of money M, in which the young generation of period t invests its saving in order to consume when old in period $t+1$. Let x_t be consumption of this generation in period t and y_{t+1} be planned consumption for period $t+1$, knowing that the nominal price of the good is p_t in period t and expected to be p^e_{t+1} in period $t+1$ (a point expectation in this deterministic framework). For simplicity let us assume that the generation has an endowment e in period t, but no endowment in period $t+1$ and that the utility of the representative individual is separable:

$$U = u(x_t) + v(y_{t+1}). \tag{112}$$

The constraints are:

$$p_t(e - x_t) = M = p^e_{t+1} y_{t+1}. \tag{113}$$

From the first-order condition for maximization (in the choice of x_t, y_{t+1} and M) we derive the optimal saving. More precisely

$$p^e_{t+1} u'(x_t) - p_t v'(y_{t+1}) = 0. \tag{114}$$

This equation, together with equality of the two extreme members in (113), show that the optimal saving is a function of the ratio p_t/p^e_{t+1}, i.e. of the price p_t for a given expected price p^e_{t+1}:

$$e - x_t = \sigma(p^e_{t+1}/p_t). \tag{115}$$

Let us call σ the saving supply function.

The temporary equilibrium of period t determines the price p_t, for a given price expectation p^e_{t+1}, either by the condition that the value of the saving of the young generation is equal to the money supply M (the first equation in (113) and (115) are solved), or (Walras identity) by the condition that the total consumption of the good $x_t + y_t$ is equal to the supply e, the consumption y_t of the old being given by $p_t y_t = M$, hence $\sigma(p^e_{t+1}/p_t) = M/p_t$.

[95] The presentation of the model is here short. The reader can find more details and discussion concerning the same model in O. Blanchard and S. Fischer, *Lectures on Macroeconomics* (MIT Press, 1989), Chapter 5, Section 5.4.

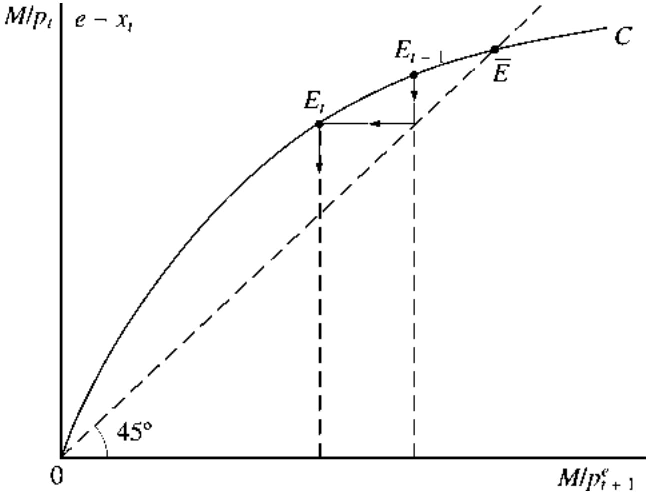

Fig. 5.

Taking advantage of equations holding in the temporary equilibrium, the present practice refers to the graph of Figure 5. The coordinates are respectively M/p^e_{t+1} and M/p_t, which is equal to $e - x_t$ because of the first equation (113). The curve C on the graph can be read (unconventionally) as giving the saving supply as a function of the slope from the origin: in the case of Figure 5, which could be shown to follow from natural hypotheses on the functions $u(x_t)$ and $v(y_{t+1})$, we may think of two possible values of this supply for a given slope p^e_{t+1}/p_t: either the origin (saving may be infinitesimal when the future expected price is tending to zero) or any other point of the curve C, E_t say. This point also gives the value of p_t in the temporary equilibrium for the given price expectation p^e_{t+1}.

This graph is convenient for a discussion of the evolution implied by the hypothesis of rational, hence exact, expectations. Indeed, the abscissa of the temporary equilibrium E_{t-1} of period $t-1$ has to be equal to the ordinate of E_t. Starting from E_t and going backward in time we then see how the antecedent E_{t-1} to E_t can be found thanks to the first bisector of the graph. Stationary equilibria have so to belong to both curve C and the first bisector. Figure 5 corresponds to the case of two stationary equilibria, the origin 0 and the intersection $\overline{E}$. We immediately see that the origin is a stable stationary exact expectation equilibrium, whereas $\overline{E}$ is unstable. We shall not discuss the

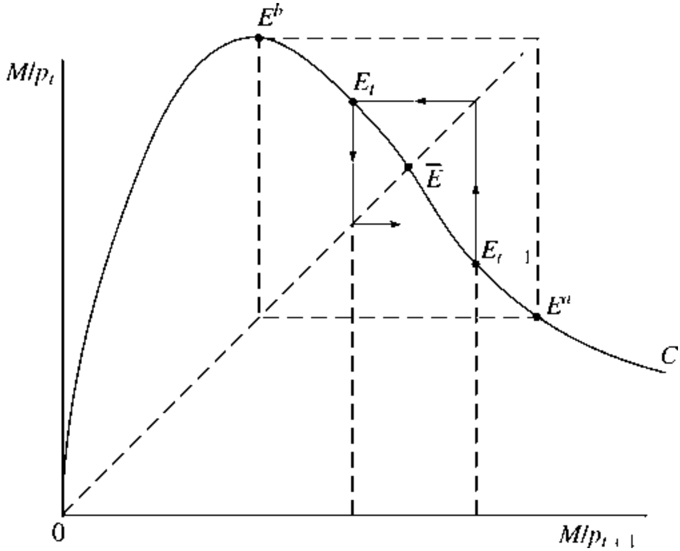

Fig. 6.

meaning of this instability because that would take us too far afield. Actually the literature on endogenous cycles focuses its attention on a different case, the one exhibited by Figure 6.

With the curve C of Figure 6, which has at $\overline{E}$ a slope lower than -1, we immediately see that the stationary equilibrium $\overline{E}$ is stable. We also see on the figure that a "cycle" exists in which the temporary equilibrium indefinitely alternates from E^a to E^b and back to E^a. More complex figures may be drawn, introducing the possibility of more complex evolutions of the temporary equilibrium, such as cycles involving three (E^a, E^b, E^c) or more points, or even a chaos (an infinite number of points with jumps leading to no kind of convergence). Some such evolutions appear stable, others unstable[96].

We shall not proceed to a discussion concerning how a path following the cycle compares with a path staying for ever at the stationary equilibrium $\overline{E}$ in terms of Pareto efficiency. Let us simply record that cases can be found in which a stable cycling path is Pareto dominated by the stationary path. If, as in Figure 6, $p^a < \bar{p} < p^b$, those have to be cases in which in the first period

[96] For a detailed study see J.-M. Grandmont, On endogenous competitive business cycles, *Econometrica* **53** (1985) 995–1046.

($t = 1$) the price is equal to p^b, so that the individual of the old generation consumes less than if the price p_1 had been equal to $\bar{p}$. Moreover all following generations are also worse off, in the sense that the consumption bundle ($\bar{x}$, $\bar{y} = e - \bar{x}$) is preferred to both (x^a, $e - x^b$) and to (x^b, $e - x^a$), where, for instance ($x^a, e - x^b$) is the consumption bundle chosen by the individual of generation t when confronted to $p_t = p^a$ and $p_{t+1} = p^b$.

Can a realistic theory of endogenous cycles be built on the basis of a cycling path of this model? We may think that we are still far from it. We do not see much in reality that would conform to the macroeconomic vision of alternating successive equilibria. Adopting the viewpoint of "microeconomic foundations", Blanchard and Fischer (op. cit.) discuss the conditions required for the two functions $u(x)$ and $v(y)$ to lead to a graph of curve C with the shape exhibited on Figure 6; they find that these conditions are quite stringent.

However, proponents of endogenous-cycle theory stress the idea that the possibility of complex dynamics, which is here discovered with a simple overlapping-generations model, has a good chance to be also present in the real world, which is so much more complex; this real possibility, they say, is completely hidden in the standard RBC model. In other words, proponents do not pretend their models to be descriptive of the macroeconomy, but rather to be revealing of a major difficulty likely to affect the coordination assumed in rational-expectations equilibria.

We leave the question at this point. We shall come back to it at the end of the section. For the time being we must take a brief look at the results obtained by the endogenous-cycle school with models that are closer to those commonly studied by the RBC movement.

In the simple one-sector one-good growth model with strict concavity of the production and utility functions the competitive equilibrium path uniformly converges to a limit path, which is the same for all values of the initial capital, as was suggested by Sections 3.2 and 3.4 of Chapter 5. But in multi-sector models difficulties can arise for two reasons, either because, with several capital goods, the technology may not imply strict concavity, or because the factor γ discounting utilities of the representative consumer may be too low. It is then possible to build examples in which the growth path will exhibit almost any sort of dynamic behaviour that will have been selected. However, Guesnerie and Woodford report that "known examples of an unstable steady state together with a reasonable [, hence not quite low,] discount factor involve asymptotically periodic, rather than chaotic, dynamics".

As long as they remain within the domain of the perfect and perfectly competitive economy, examples of cyclical or other atypical evolutions do not much challenge the neoclassical conception of the world: strictly speaking they do not dispute Pareto-efficiency; they might simply lead to the suspicion that small deviations from the perfect economy might seriously damage efficiency. This relative unimportance may explain why, in order to find atypical dynamics and endogenous cycles, economists are often now looking outside the domain in question, in directions where pathologies seem likely to be more frequent and more serious. Among the various market imperfections, some were particularly examined in the literature now discussed. Overlapping-generations models consider a case in which the market system is viewed as incomplete or the access of agents to markets is viewed as limited.

Independently of this case, liquidity constraints are often envisaged. Their presence may lead to the more or less periodic occurrence of large jumps; for instance, in a two-sector model similar to the one studied in Section 5.1 of Chapter 6, but with less strictly determined saving rates, the possibility exists for cycles of the same type as the one there described; but under different hypotheses chaotic evolutions can be found even in cases where the discount factor γ is high.

Attention was also brought to a case that must now appear familiar to readers of this book, the case in which increasing returns to scale and monopolistic competition coexist. Another interesting case has a technology with economies of scale that are external to the firms; the model can then draw advantage from the formalization defined at the end of Section 6.2 of Chapter 4 and used in Section 7.2 of Chapter 6; but users of the model can also consider the growth paths obtained when the technology does not lend itself to the easy aggregation used in those sections. Accommodated into an equilibrium framework, either through monopolistic competition or through externalities, increasing returns to scale make again easier the appearance of unconventional dynamics.

2. Let us now turn attention to *the possible occurrence of sunspot equilibria* and to their role in stochastic macrodynamics. Thus far in this book, the presence of randomness in the economic system was traced to shocks concerning the fundamental elements of the model describing this system. In particular proponents of the RBC approach insist that these shocks have to concern technology or preferences, or still endowments of individual agents, unless a public sector is also considered. It is common to say in the literature now to be discussed that such uncertainty is "intrinsic": it is embodied in the fundamentals of the system.

"Not all economic randomness can be explained in this way. Even if the fundamental parameters were non-random, economic outcomes would generally be random. This is because the economy is a social system composed of individual economic actors who are uncertain about each other's behaviour. In seeking to optimize his own actions, each participant in the market economy must attempt to predict the actions of other participants... Since market participants are not certain about the actions of others, they are uncertain about economic outcome"[97]. It then so happens that equilibria exist in which outcomes are random even though no randomness exists in fundamentals. Agents may explain this randomness by sunspots if they wish; but sunspots actually have no effect on the economy, except through agents' beliefs. The now familiar denomination "sunspot equilibrium" keeps, out of context, reference to this hypothetical case about the rationalization of beliefs.

Uncertainty of this sort, which is not transmitted through the fundamentals may be called "extrinsic uncertainty" or, probably better, "market uncertainty" (the language is a bit confused when it must convey the idea that "endogenous cycles" may be due to "extrinsic uncertainty").

"The interdependence of beliefs, even of rational beliefs, is a central theme in the *General Theory* (Chapter 12). Keynes postulates that it is possible to encounter self-justifying expectations, beliefs which are individually rational but which may lead to socially irrational outcomes"... "Indeed, research on sunspot equilibrium was inspired by and in reaction to the rational-expectations macroeconomics literature... Most of the critics took issue with the assumptions of individual rationality and perfect markets... Others... were willing to ask whether or not the conclusions of the rational-expectations school follow from the assumptions." (again quotes from K. Shell, op. cit.).

In order to illustrate what sunspot equilibria mean, an example will be given in the overlapping-generations model examined a moment ago. There is no intrinsic uncertainty in the sense that all elements of the model are assumed to be known for sure. However, if agents believe in the existence of a sunspot stochastic process affecting the equilibrium path, this belief may be confirmed; in that case, expectations will be fulfilled, hence rational.

Let the sunspot process be a two-states Markow process, with a high state s^h and a low state s^l. When the high state holds in period t, there is a probability q^h that s^h will again prevail in period $t+1$ and a probability $1-q^h$ that $s_{t+1}=s^l$. Similarly q^l is the probability that the state remains at s^l in the period just following one in which it was already there. Let us limit attention to stationary but stochastic equilibrium paths. In conformity with the existence of the sunspot process agents expect in period t that the price p_{t+1} will be equal to p^h or p^l depending on whether $s_{t+1}=s^h$ or $s_{t+1}=s^l$. So when s^h prevails in period t, the representative individual of generation t maximizes:

$$u\left(e-M/p^h\right)+q^h v\left(M/p^h\right)+\left(1-q^h\right)v\left(M/p^l\right). \tag{116}$$

[97] Drawn from K. Shell, Sunspot equilibrium, entry in: J. Eatwell, M. Milgate and P. Newman, *The New Palgrave Dictionary of Economics* (Macmillan, London, 1987).

The first order condition then is

$$\frac{1}{p^h}u'\left(e - M/p^h\right) = \frac{q^h}{p^h}v'\left(M/p^h\right) + \frac{1-q^h}{p^l}v'\left(M/p^l\right). \tag{117}$$

And by symmetry the equation where p^h and p^l are interchanged, q^l replacing q^h, also holds. For simplicity let us write $u'(h)$ instead of $u'(e - M/p^h)$ and $v'(h)$ instead of $v'(M/p^h)$. Then Equation (117) may be written in the form:

$$\frac{q^h}{1-q^h} = \frac{p^h v'(l) - p^l u'(h)}{p^l u'(h) - p^l v'(h)}. \tag{118}$$

Symmetrically the first order conditions also require:

$$\frac{q^l}{1-q^l} = \frac{p^l v'(h) - p^h u'(l)}{p^h u'(l) - p^h v'(l)}. \tag{119}$$

Equations (118) and (119) might be seen as proving that, p^h and p^l being arbitrarily chosen, there is a sunspot stationary equilibrium process with the price level randomly shifting between these two prices p^h and p^l, in conformity with the shifts of the corresponding (q^h, q^l) Markov process between the two states s^h and s^l. But this is not quite true: indeed it is true if and only if, given p^h and p^l, the right-hand sides of (118) and (119) are both positive, so that q^h and q^l are both positive and smaller than 1 (remember moreover that equilibrium values $u'(h)$, $u'(l)$, $v'(h)$ and $v'(l)$ depend on p^h and p^l). A precise mathematical discussion of this condition would take us too far.

Let us simply quote the following two results applying to a search for sunspot equilibrium processes in which $p^h - \bar{p}$ and $p^l - \bar{p}$ would be infinitesimal, $\bar{p}$ being the price level in the stationary deterministic equilibrium $\overline{E}$ of either Figure 5 or Figure 6: (i) in the case of Figure 5, whatever the choice of p^h and p^l, there is no such sunspot stationary equilibrium process, (ii) in the case of Figure 6, there is a range of values of (p^h, p^l) for which sunspot stationary equilibrium processes exist. (For the proof, see O. Blanchard and S. Fischer, op. cit.).

We shall not repeat the cautious remark made about the realism of the endogenous cycle exhibited by Figure 6, a remark that would now apply to the realism of the sunspot equilibria. We shall rather draw from the survey of Guesnerie and Woodford a few indications about existence of sunspot equilibria in other dynamic models than the one to which Figures 5 or 6 applies. (Examples of sunspot equilibria in static models also exist.)

Sunspot equilibria can be found with models in which the individuals are living for ever and have to invest in a flat money, either because of a cash-in-advance constraint for certain goods or because they cannot borrow against future income. Sunspot equilibria are also known to be possible in a variety of non-monetary dynamic models, which may involve a finite number of infinitely-lived consumers rather than overlapping generations. For instance, the presence of short-run increasing returns to scale may make sunspot equilibria possible, either with monopolistic competition or even with perfect competition if the increasing returns are external to the firms as in Section 6.2 of Chapter 4. Even if returns to scale are not increasing in the short run, the possibility of longer-run sunspot equilibria may result from the complementarity between different types of capital goods, whose invention and development have a fixed cost financed by temporary monopoly rents[98]. Finally, examples of sunspot paths can also be given in models with sticky prices and rationing.

3. As was reported at the beginning of this section, it is still difficult to judge *how the theoretical possibility of endogenous cycles and of pure market uncertainty should affect macroeconomic teaching*. Advocates of the relevance of this theoretical possibility may argue, alternatively or simultaneously, on three grounds.

First, they may focus on the microeconomic foundations of macroeconomics. They may stress the idea that perfect Arrow–Debreu economies give an inappropriate foundation for discussing the macroeconomic instability that is so obvious to observers. They may explain in general terms how even market clearing and rational expectations are compatible with inefficient instability. They may request from proponents of "the new neoclassical macroeconomics" that they be more moderate in their methodological claims. This line of argument is, of course, valid but not constructive. Some macroeconomists may think that demonstration of the limits of the Arrow–Debreu world has been so well given that it is now urgent to concentrate on more positive work about macroeconomic problems.

Second, advocates of the theory of endogenous cycles may point to the importance of sharpening our judgement about the effects of the range of phenomena which are alleged by various authors to be responsible for endogenous cycles. They may argue that macroeconomists have to appreciate which theoretical analyses they find most convincing; macroeconomists therefore need to be provided with models proposing precise treatments of alternative views about the actual nature and role of endogenous cycles. The recent research,

[98] See G. Evans, S. Honkapohja and P. Romer, Growth cycles, *American Economic Review* (June 1998).

advocates may say, is doing exactly that: the outcome may not appear immediately useful, but contributions of this research will mature and progressively lead to positive results which will serve macroeconomists. In other words, the theory of endogenous cycles would still be in its infancy, but its relevance would be bound to emerge with time.

If we speculate about such a prediction, we have to accept the idea that, for still some time to come, macroeconomists will have to make their own selection among the models exhibiting alternative theories of endogenous cycles. They will then depend on their *a priori* knowledge or beliefs concerning the respective realism of various sets of hypotheses[99]. We may feel reluctant at basing judgement on the rather diffuse notion of *a priori* knowledge, but we have to recognize that, often in the past, such was the way through which our science developed.

If the reader accepts the prediction in question, he or she would probably like to find in this book a representative sample of the models now available, so as to be able to make the investment of closely studying the sample and to form his or her judgement. But that exceeds what this book can do, because well surveying a number of difficult analytical works would require too much space.

The author can, however, express his own main provisional conclusion, which is neither specific nor probably surprising. The main positive outcome of the current research is likely to concern the dynamics of business sentiments: the "animal spirits", whose role was identified by Keynes in the *General Theory*, the optimism or pessimism of households, the confidence in the success of government actions, and so on. Understanding the spontaneous development of such business sentiments, interacting with the economic evolution permitted by technological, institutional and market structures, would be a significant achievement. Within the treatment of such a topic self-fulling expectations would play a role, but it would be fairly easy to leave also room for the permanent or occasional deviations to which we refer when we speak of bounded rationality.

Third, we begin to see some research aiming at showing that models of endogenous cycles fit macroeconomic observation. In particular, students can now benefit from the publication in June 1994 in the *Journal of Economic Theory* of a symposium on "Growth, fluctuations and sunspots: confronting the data".

[99] In order to reflect about this dependence, the reader may consider the article by Evans, Honkapohja and Romer (op. cit.). The authors claim that the crucial assumptions of their model are selected because of their plausibility, a quality which they do not grant to assumptions made in other models of endogenous cycles. Will the reader be persuaded?

More precisely we may consider here the article by E. Farmer and J.-T. Guo[100], not because the application would be fully convincing, but because it takes seriously the investigations to be led for any model claiming to have descriptive value; at least it takes them as seriously as, even maybe somewhat more seriously than, do common RBC applications. The upshot of the article is to compare the empirical performance of a standard RBC model to that of a model with no internal shocks to productivity but with self-fulling beliefs conveyed by "animal spirits".

The new model has increasing returns to scale and monopolistic competition; the "sunspot process" representing agents beliefs is a white noise process. The authors well show the difficulty of calibration of some parameters and explain their choices, which may not all be easily accepted by readers (for instance, the assumption that the labour supply is infinitely elastic with respect to the real wage rate; but the same assumption is also made in the calibration of the RBC model taken for comparison). The degree of increase in returns to scale is rather high, as well as the microeconomic degree of monopoly power; it follows that the demand for labour may be characterized by an increasing function of the real wage rate. The authors explain why such high values of the two parameters are necessary for the existence of sunspot dynamic equilibria and why they believe that such values belong to the domain of reasonable estimates for these parameters.

The empirical fit to quarterly US data for 1954 to 1991 is found to be somewhat better as a whole with the sunspot model than with the RBC model. For relative volatilities and contemporaneous correlations with output the performance is roughly the same: the sunspot model captures better volatilities of investment and labour input, but is worse on volatilities of consumption and productivity. The main difference comes from the fact that the sunspot model is rather successful in tracing dynamic patterns without building them into the driving process, since sunspots are assumed to behave like a white noise whereas the RBC model assumes highly autocorrelated productivity disturbances: with the new model the Solow residuals broadly resemble those in the US time series; the response to a unit innovation in output is cyclical, as it is in a VAR fitted on the data, whereas the RBC model predicts a monotone return to equilibrium.

We cannot yet draw general conclusions from this work. It will have to be duplicated on other data sets and more thoroughly discussed. But the attempt is challenging.

[100] E. Farmer and J.-T. Guo, Real business cycles and the animal spirits hypothesis, *Journal of Economic Theory* **63** (1994) 42–72.

2.7. The inventory cycle

During four sections the focus was on general equilibrium. We now move to more analytical studies taking in turn various aspects which, time and time again, revealed to be particularly important in business fluctuations. A common feature of these various aspects is to be closely related, for one reason or another, to price and wage stickiness.

Discussing in Section 2.2 stylized facts of the economic cycle, we mentioned that business inventory investment was recognized as being strongly procyclical and as being even leading output. But such a description of the relation between changes in inventories and the business cycle is not complete. It applies the common approach to the measurement of stylized business cycle facts. But it overlooks something else, which also seems to be systematic in the relation between the inventory cycle and the output cycle. The explanation requires that we first question the image which, more or less explicitly, provided the reference for our study of economic fluctuations up to now in this chapter, namely that of a regular long-run trend of output on which are superimposed not only erratic variations but also smooth fluctuations with profiles similar to those exhibited by a sine function of time.

After bringing attention to stylized facts, we shall survey models which, dealing with the role of inventory investment in the business cycle, will often have to distinguish between a passive and an active roles, changes in inventories being partly involuntary and partly deliberate (see Section 5.5 of Chapter 4). In Subsection 2 we shall consider a model in which inventory accumulation will be just a latent variable but output fluctuations will be explained by inventory behaviour interacting with supply and demand shocks. Specifications examined in Subsection 3 will on the contrary identify inventory accumulation; they will then be related to microeconomic investigations.

1. The idea is more and more widely accepted that, beyond a good deal of variation from one cycle to another, *a systematic asymmetry* tends to prevail: in the cyclical component, depression phases are shorter than expansion phases[101]. A recent article[102], from which we shall borrow here, even argues that the typical business cycle has three phases: the upturn is followed by a relatively short period of high growth (the "bounceback phase"); before the third phase (the depression), the second, more or less long, phase is characterized by

[101] S. Neftci, Are economic time series asymmetric over the business cycle?, *Journal of Political Economy* **92** (1984) 307–328.

[102] D. Sichel, Inventories and the three phases of the business cycle, *Journal of Business and Economic Statistics* **12** (1994) 269–277.

moderate growth. This corresponds to an image once given by Milton Friedman[103]: "output [should be] viewed as bumping along the ceiling of maximum feasible output except that every now and then it is plucked down by a cyclical contraction... When subsequent recovery sets in, it tends, to return output to the ceiling".

In order to establish this result, Sichel applies various statistical techniques to the quarterly series of the US real GDP growth rate from 1950 to 1992. In particular, he shows a chart exhibiting average growth rates computed after merging quarters similarly located with respect to the date of the upturn, as determined by the National Bureau of Economic Research (NBER). These rates go down from over 6 percent (on annual base), during the two first quarters after the upturn, up to something like 3.5 percent, starting after the sixth quarter. In order to avoid reference to the identification of turning points, Sichel also establishes the significance of the "output gap" as an additional regressor in an autoregression of the growth rate of real GDP[104]. The same existence of three phases appears with a quarterly series of an index of US industrial production from 1884 to 1940.

This vision of three phases tends to invalidate methods of linear dynamic analysis, which prevailed in economic teaching about business fluctuations (including in this book). Sichel argues that it even invalidates econometric results about the long-run trend seen as the realization of a stochastic process (see below). This vision also invalidates the non-linear accelerator-multiplier theory of business cycles which was proposed by John Hicks and discussed here in Section 1.3: as shown by Figure 2, the theory in question would imply that sharp changes in the pace of growth would occur at downturns rather than at upturns. Understanding of the phenomenon is, all the more, required.

Sichel argues that the source of the three-phase pattern is found in inventory investment. Indeed, the remaining part of aggregate demand, which may be called "real final sales", does not exhibit the bounceback phase. This is apparent on Table 1 distinguishing the contributions to GDP growth of respectively final sales and inventory investment. The table distinguishes two phases within expansions: the first year, the following expansion. The contribution of real final sales reaches, on average over all cycles, the same level in the two phases. But inventory changes dominate in the depression, bring a still more important absolute contribution to growth in the first year after the upturn, but no longer contributes in the following expansion.

[103] M. Friedman, *The Optimum Quantity of Money and Other Essays* (Aldine, Chicago, 1969), p. 274).

[104] The output gap is measured as the log of the ratio between the last maximum value of real GDP and its current value. It is introduced with a one-quarter lag in the auto-regression.

Table 1
Contribution of inventory investment in the three phases of the cycle (annual growth rate, US, 1950–1992)

Contributions	First year after upturn	Following expansion	Depression
Final sales	4.1	4.0	−0.7
Inventories	1.7	0.0	−1.4
GDP (Total)	5.8	4.0	−2.1

Source: D. Sichel, op. cit.

We saw previously, first in Section 5.4 of Chapter 8, and second in Section 2.1 of this chapter, the importance of a proper representation of the long-run trend in stochastic approaches to macroeconomic dynamics. The main question was to know whether this trend was subject to significant and recurrent permanent shocks like in a random walk. That something of the sort exists is commonly accepted today. But there is a risk that too much weight could be given to permanent shocks in comparison with transitory shocks, which concern the short-run dynamics. The risk may materialize if macroeconomic time series are processed in econometric analysis as if they were realizations of ARIMA processes whereas the true underlying phenomenon is better described by a non-linear process. Implications of such a misspecification in the US real GDP analysis were indeed discussed in Section 2.1.

Since swings in inventory investment account for the lion's share in the variance of output growth, a natural question in this discussion is to know whether they too exhibit the presence of a permanent component. Examination of the question by Sichel using the standard Dickey–Fuller tests led to a strong rejection of the hypothesis that the process of inventory investment had a unit root (the tests assume an ARIMA process, with possibly a linear deterministic trend; but the rejection is so strong that it seems likely to hold also with less linear maintained hypotheses).

2. Considering the importance of the inventory cycle in descriptions of business fluctuations, we understand that some theorists are looking for *dynamic models in which inventory behaviour would play the strategic part*. Such is the case of Guy Laroque, the research of whom will now be examined[105].

[105] G. Laroque, The inventory cycle and the instability of the competitive mechanism, *Econometrica* **57** (1989) 911–936.

The initial idea derives from the facts that, first, the two categories of nominal assets on the one hand, and physical goods held in inventories on the other, are close substitutes as reserves of value, and that, second, their respective returns may substantially differ and substantially vary from one period to the next (with close substitutes such could not be the case if market prices were perfectly flexible). Portfolio decisions then lead at times to substitute nominal assets to stocks of physical goods or the reverse, depending on what are the expectations about the respective future returns. This speculative behaviour explains why changes in inventories do not come only from unanticipated changes in transaction needs or from changes in expected future such needs.

As we saw in Subsection 1.6.2, the terms of the arbitrage may be given by either the expected rate of price increases or the expected real rate of interest, depending on which nominal assets provide the relevant substitute to inventory holding, either money or interest-bearing financial assets. If the nominal interest rate does not change, the two indicators are quantitatively equivalent. As long as we focus on cases of rapid changes in expectations about inflation, the two indicators remain qualitatively equivalent because changes in nominal interest rates then seldom fully reflect changes in those expectations. Be that as it may, Laroque studies the case when the terms of the arbitrage are given by the expected rate of price increases.

Considering moreover the case of sluggish price adjustments and of rationing schemes opening opportunities to the realization of speculative gains, Laroque (op. cit.) shows that many trajectories are feasible for dynamic equilibria with self-fulfilling expectations. Most of these trajectories are made up of successive phases. Let us first consider phases of "normal times". That may be a phase of optimistic speculative behaviour: rapid price increases are expected; inventory investment is important and boosts aggregate demand; actual prices therefore increase fast; anticipations are validated. There are also phases of normal times with pessimistic speculative behaviour, strict management of inventories maintained to the minimum, depressed aggregate demand, therefore self-fulfilling deflationary expectations. In principle phases with neutral behaviour and exact market clearing are feasible. But in the model built by Laroque, phases of this third type are unstable, whereas phases of the two other types, although not likely to be indefinitely sustained, are stable with respect to the formation of the temporary equilibria. In contrast instability of the competitive stationary equilibrium is a crucial feature of this theory. When normal times of inflation end, expectations are abruptly revised, inventories quickly decrease and activity falls down much below capacities. After times of deflation recovery is more gradual, even though there is a sharp discontinuity in the evolution of expectations, of excess demand and of the rate of inflation.

This theory is clearly related to the one conveyed by the models of endogenous cycles discussed in the previous Section 2.6. It even more precisely concerns the last subsection in which we wondered whether the theoretical possibility of endogenous cycles had relevance for macroeconomists studying actual business cycles. But can the theory be tested?

In an attempt to answer the question G. Laroque and G. Rabault[106] had first to build an econometric model that would well correspond to the theoretical model. They reasoned as follows. Two regimes had to be distinguished, each period t belonging to one or the other: either there was excess demand, output was equal to supply and rapid price increases were expected ($s_t = +1$), or there was excess supply, output was equal to demand and slow price increases were expected ($s_t = -1$). Considering the logarithms for all variables, their notation is y_t for output, y_t^* for supply and y_t^d for final demand excluding inventory investment, hence:

$$y_t = \begin{cases} y_t^* & \text{if } s_t = 1, \\ y_t^d & \text{if } s_t = -1. \end{cases} \tag{120}$$

Notice here the following assumptions: in the regime of excess demand, voluntary inventory investment is assumed to be positive (see below), so that aggregate demand is larger than final demand and is rationed; involuntary inventory investment then is negative; on the contrary in the regime of excess supply, both voluntary and involuntary inventory investment are nil (again see below).

Supply is assumed to be ruled by an exogenous random walk with drift:

$$y_t^* = y_{t-1}^* + \mu^* + \varepsilon_t, \tag{121}$$

where $\{\varepsilon_t\}$ is a pure Gaussian white noise process. Final demand is also assumed to be ruled by a random walk, but with a drift that depends on the regime according to:

$$y_t^d = y_{t-1}^d + \mu(s_{t-1}) + \eta_t. \tag{122}$$

Like as the supply shock ε_t, the demand shock η_t is a Gaussian white noise.

Moreover the following inequalities are assumed:

$$\mu(1) < \mu^* < \mu(-1). \tag{123}$$

[106] G. Laroque and G. Rabault, The inventory cycle: from theory to empirical evidence, *Economic Journal* (March 1995).

The rationale for the assumption, as well as for those made above about voluntary inventory investment, is found in movements of the expected price level, a variable which does not appear explicitly in the model: this expected price level is assumed to increase fast in the regime of excess demand ($s_t = 1$) and slowly in the regime of excess supply ($s_t = -1$), then at a rate smaller than the rate of return on other uses of funds than inventory accumulation. So, in the regime of excess demand: (i) inventory investment is more rewarding than alternative uses of funds, (ii) capital losses are incurred on money and nominal assets, so that final demand grows slowly. In the regime of excess supply on the contrary: (i) inventory investment is less rewarding than alternative uses of funds, (ii) capital gains are realized, so that final demand grows fast.

A similar rationale explains the assumption made about the stochastic process followed by the state variable s_t. If excess demand prevails in period t ($s_t = 1$), private agents will aim at realizing voluntary inventory investment in period $t + 1$, as long as prices will keep increasing fast. The authors assume that, in that case, the logarithm of aggregate demand will be $y^d_{t+1} + s$, s being an exogenous positive number. They then write:

$$\begin{aligned} &\text{if } s_t = 1 \text{ and } y^d_{t+1} + s > y^*_{t+1} \quad \text{then } s_{t+1} = 1, \\ &\text{if } s_t = 1 \text{ and } y^d_{t+1} + s \leqslant y^*_{t+1} \quad \text{then } s_{t+1} = -1. \end{aligned} \tag{124}$$

On the contrary, if excess supply prevails in period t, private agents will aim at a zero inventory investment, hence:

$$\begin{aligned} &\text{if } s_t = -1 \text{ and } y^d_{t+1} < y^*_{t+1} \quad \text{then } s_{t+1} = -1, \\ &\text{if } s_t = -1 \text{ and } y^d_{t+1} \geqslant y^*_{t+1} \quad \text{then } s_{t+1} = 1. \end{aligned} \tag{125}$$

We note that the transitions are no longer ruled by exogenous probabilities, as they were with Equations (72) posed by J. Hamilton in his non-linear specification of the process of GNP growth; they now depend on the realizations of the stochastic processes $\{y^*_t\}$ and $\{y^d_t\}$. More importantly, we must now reflect on the extra assumption, which is so embodied in (124)–(125), namely that private agents behave in such a way as to remain in the same regime as long as they do not see evidence showing a change of regime. This a natural assumption, and we cannot think of any other sensible and better assumption for the determination of the process $\{s_t\}$: assuming it to be just an exogenous stochastic process would be denying that inventory behaviour played a strategic part in the business cycle.

In order to test their model and the theory underlying it, Laroque and Rabault faced a real challenge: indeed, only the series of output y_t was taken as observed, s_t, y_t^* and y_t^d being just latent variables. The challenge came from the paucity of the information on which inference had to be based. The authors used the fairly long quarterly US GNP series from 1947 to 1992. The challenge had an economic side: against which alternative model should the performance of the proposed model be tested? We shall shortly come back to the question. We want first to recognize that the challenge had also an econometric side: it was not so easy to determine the method thanks to which all the unobserved elements of the proposed model could be found.

The authors made a first exploration. They used the NBER business cycle chronology and the GNP series for a calibration of values to be given to the parameters: μ^*, $\mu(1)$, $\mu(-1)$, s and the standard deviations of the Gaussian errors ε_t and η_t. They then ran a number of simulations (1.000) with their model so calibrated, in order to proceed to a qualitative comparison of these simulations with the GNP series. Some features were satisfactory, such as an average cycle length of four years. But the observed series is much smoother than the simulations so obtained. This is easily traced to the fact that the model defined by (120) to (125) admits not enough serial correlation when the unit period is understood to be a quarter.

Knowing that a good fit to the quarterly GNP series is obtained with a trend and an autoregressive process of order 2, the authors revised their model, introducing first- and second-order autoregressive terms in the demand and supply processes. They then worked out and applied an econometric procedure for the estimation of the model.

As for the test of the model, the natural comparison was with a simple second-order linear autoregression. It then turned out that the data were not informative enough to statistically discriminate between the two specifications. However, the investigation led to encouraging results. They concerned first the fit to the data and second the type of business cycle dynamics that the proposed model implied.

An interesting by-product of the statistical procedure was an estimate of the probability density functions of the latent variables (supply, demand and regime), conditional on the current and past observations of GNP. In particular the estimated probabilities of being in the excess supply regime were always very low at NBER peaks, which were all followed by sharp increase in this probability; but some peaks in the probability also occurred twice at times which had not been identified as special in the NBER chronology.

A thousand simulations of 180 quarters series were computed with the estimated model. They resulted in a variable number of cycles: 6.3 on average, but with a standard deviation of 2.3. This agrees with the stylized fact that most cycles have length varying from 5 to 10 years (Section 2.2, Subsection 5). On average, phases of excess supply lasted for 7.7 quarters, phases of excess demand lasted for 22.3 quarters, well portraying the asymmetry identified by Sichel.

3. *Considering directly data on inventories and on their immediate determinants* amounts to look at the same phenomenon, but from the opposite point of view: no longer as a structural latent explanation leading to a univariate model of observed changes in GNP, but as an observable block in the building of a more embracing theory to be elsewhere considered. We already adopted this second and familiar point of view in Section 5.5 of Chapter 4. But, although hopefully useful for a first examination of inventory behaviour, the section did not give a complete account of what should be said now in a business cycle perspective. In order to complement what was written in Chapter 4 our reference will be an article by A. Blinder, who earlier devoted a good deal of research on inventory behaviour, and L. Maccini[107].

The authors first point to stylized facts concerning the US economy but certainly applying to all developed economies, if not even more generally. They first note that manufacturers' finished goods inventories play a relatively minor part in business fluctuations: between 1959 and 1986 they amounted to 13 per cent of inventory investment, whereas investment in manufacturers' inventories of "materials and supplies" amounted to 27 per cent and inventory investment in trades to 50 per cent. The variance of the monthly inventory investment in finished goods in manufacturing was only 6 per cent of the variance of aggregate investment in inventories.

The observation is important because by far the largest part of the literature referred to, or concerned, a model inspired by the idea of production smoothing: stocks would act as a buffer somewhat insulating production from variations in sales. Although we did not explicitly pose the model in Section 5.5 of Chapter 4, we referred to it and explained how it could well account for the fact that there was actually no production smoothing: time series of output and sales exhibit a higher variance of output than of sales. Two rationalizations were mentioned: there is an acceleration phenomenon (since the volume of inventories has to be roughly proportional to sales, an increase in sales requires an extra output for inventory investment); fluctuations in costs induce fluctuations in profit-maximizing output in addition to what fluctuations in sales may require. There is much more to say about the implications and econometric ap-

[107] A. Blinder and L. Maccini, Taking stock: a critical assessment of recent research on inventories, *Journal of Economic Perspectives* (Winter 1991).

plications of the production-smoothing model[108]. But, since the model cannot pretend to be obviously appropriate beyond the case of manufacturers' finished goods which play a minor part in business fluctuations, we must pay attention to the existence of another model which has long been used for rational inventory management.

This alternative model focuses on inventories in commercial firms and inventories of materials and supplies in industry, construction or agriculture. It formalizes the management of a stock from which drawings are made when required for sales or for use in production, a stock which is periodically replenished by purchases, each time with a delivery lag and a fixed cost coming in addition to the value of the purchase, possibly also with a penalty in case of stock-out. In a deterministic or stochastic stationary environment the optimal policy of the manager is defined by two numbers, s and S, and by the rule to order a purchase as soon as the volume of the stock falls at or below s and then to buy such an amount as to replenish the stock to S. We actually referred to such a case and to such a policy when we discussed the demand for money resulting from the transaction and precautionary motives (Sections 4.2 and 4.3 of Chapter 2).

What Blinder and Maccini call "the (S, s) model" assumes that inventory holders apply the above policy, each one with its own numbers s and S, which remain fixed over the period under consideration. By definition of the policy, inventory investment is discontinuous at the microeconomic level, so that for instance in wholesale or retail trade the correlation over time of inventory investment with the much more continuous sales is small; it is small but positive. This last feature is precisely what is observed at the macroeconomic level and what is responsible for the fact that output has a larger variance over time than sales.

Reflecting on this alternative model leads us to think that it does not lend itself easily to aggregation. Aggregate orders resulting from various levels of aggregate sales depend on the distribution of inventories across firms: if the distribution happens to be skewed toward s, the reaction will be weak. However, Blinder and Maccini explain how this aggregation difficulty was solved in various research approaches. They also discuss the difference between the predictions of the resulting aggregate model and those of the more common production-smoothing model. Although an equation similar to Equation (275)

[108] On these implications and applications, see the references given in Section 5.5 of Chapter 4, particularly to an article published by A. Blinder in 1986. See also pages 78 to 85 in the 1991 article of Blinder and Maccini. For an apparently still more complete treatment see K. West, Inventory models: the estimation of Euler equations, in: H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics-Macroeconomics* (Blackwell, Oxford, 1995), pp. 188–220.

of Chapter 4 may be proved to hold as an approximation, the coefficient λ has an entirely different interpretation: it is no longer a measure of the speed of adjustment but a product of aggregation; it so depends on the joint statistical distribution of inventories and sales across firms, a distribution that slowly adapts after any temporary sales shock.

Surveying more generally the recent macroeconomic work about inventory fluctuations, Blinder and Maccini make a number of points, two of which are interesting here as complements to what was written in Chapter 4. Starting from the good empirical performance of a simple Keynesian model of the inventory cycle, they point to the fact that the adaptive expectations hypothesis embodied in this model is not crucial because stochastic models with rational expectations were built, which behave quite similarly[109].

They also signal interesting simulation results obtained by M. Lovell[110], who built a multi-sector model in which firms hold inventories of both raw materials and finished goods. The raw materials are guided by (s, S) rules while the finished goods are guided by conventional flexible accelerators. Production is found to be more variable than final demand because of the policy applying to raw materials. Moreover, if firms adjust their trigger points (s, S) frequently due to changes in economic conditions, tremendous liquidations of stocks take place at cyclical peaks, and vice versa at cyclical troughs; so the simulations reproduce the stylized fact identified by D. Sichel (op. cit.).

If we compare the phenomena discussed in most of the literature about inventory cyclical behaviour with the phenomenon studied by G. Laroque, we note an important difference. Whereas Laroque considers the speculative behaviour of inventory holders faced with changes in price trends, the literature commonly studies how managers of inventories react to observed or expected changes in quantities sold or used. An interesting microeconomic research might aim at assessing the part played, and the form taken, by speculative behaviour in inventory behaviour.

2.8. The productivity cycle and capacity utilization

As we just saw, inventory behaviour of firms requires a particular attention when we want to better know business cycle phenomena. Other components

[109] A. Blinder and S. Fischer, Inventories, rational expectations and the business cycle, *Journal of Monetary Economics* (November 1981).

[110] M. Lovell, Simulating aggregate inventory behavior, *Journal of Economic Behavior and Organization* **21** (1993) 147–179.

of firms' behaviour similarly need to be scrutinized more deeply than is convenient in a general equilibrium model of the economy. In this spirit let us turn now to the characterization and effects of decisions concerning productive capacities and their use. We already dealt with the characterization in Chapter 4 and with the effects on the pace of inflation in Chapter 8, particularly in Sections 2.3 and 6.6. Taking in this chapter the viewpoint of business fluctuations, we shall add complements to what was earlier exposed in our book and focus first on the productivity cycle, second on cyclical variations in the consumption of intermediate goods, third on the evolution of productive capacities.

1. We introduced the concept of the productivity cycle in Chapter 4, Section 1.4, Subsection 4, where we explained it by the short-run demand for labour of firms facing an irregular demand for their output and having to bear either adjustment costs or profit losses. The rationale there presented has been the object of a direct investigation and of a special survey of US plant managers by J. Fay and J. Medoff[111].

These authors first note that it is somewhat misleading to speak of "*labour hoarding*" as is currently done, and as was done here in Chapter 4, in order to explain the decline in measured labour productivity during depressions. They write: "Some of the tasks in the production process support production but do not contribute directly to output [as usually defined] in the period during which they are performed. Such task might include routine maintenance of equipment, overhauling machinery, cleaning, painting, training, and so on... Firms can postpone some of these tasks during periods when output is high. For these firms, the amount of labor that is needed during a downturn is greater than the amount that would be needed if the lower level of output were permanent since additional labor is needed to perform the postponed work". Accordingly, Fay and Medoff introduce distinctions: labour may be employed in regular production; it may be employed in worthwhile other work; or it may be hoarded. When it is so more strictly defined, labour hoarding still takes different forms: it may concern labour employed in regular production if effort is below its normal level, whether that happens for all employees or for just a few, whether it means longer breaks or lower rhythm; it may similarly concern labour employed in worthwhile other work, whether more labour is assigned to the task than is technically necessary, or the task is a make-work of little inherent value.

In order to know both the importance of strictly defined labour hoarding and the nature of employment in worthwhile other work, Fay and Medoff made a survey using a carefully patterned and worded questionnaire, which would be

[111] J. Fay and J. Medoff, Labor and output over the business cycle: some direct evidence, *American Economic Review* (September 1985).

too long to describe here. The questionnaire was sent to a random sample of 1000 manufacturing establishments with 100 or more employees; 326 questionnaires were returned, of which 242 were properly completed and 168 were used (74 came from managers who had not witnessed a downturn in production). Various checks led to the conclusion that the sample of the 168 used questionnaires was not biased for the purpose of the investigation.

The results are so summarized: "During its most recent trough quarter, the typical plant that had a downturn paid for about 8 per cent more blue-collar labor hours than were technologically necessary to meet that quarter's regular production and operations requirements. About half of this labor could be justified by the value of other work... that was completed during the trough quarter. Thus,... about 4 per cent of the labor paid... during the trough quarter should be classified as hoarded". Fay and Medoff also give interesting output elasticities of blue-collar hours worked in manufacturing plants: the direct calculation from actual hours gives an elasticity of 0.82 with respect to output; after correction for labour hoarding the elasticity amounts to 0.96 and after the further correction for labour employed in worthwhile other work it rises to 1.17, so exhibiting a decrease in the marginal productivity of labour in the production function corresponding to a normal use of the productive capacity by blue collar workers.

These microeconomic observations thus validate the vision of the productivity cycle as due to the intertemporal behaviour of firms exposed to adjustment costs and other rigidities preventing full flexibility of labour input. However, the numerical values just reported concern only a rather small part of employment: blue-collar workers in manufacturing plants. They are unlikely to apply as such to the rest of the labour force, the nature and importance of "labour hoarding" and of "other worthwhile work" varying from an activity to another. Moreover, as we shall now see, a simple transposition from the micro to the macroeconomic level is likely to underestimate the importance of the phenomenon.

2. The underestimation in question belongs to the consequences which might follow from the *consideration of intermediate outputs and inputs* within the productive system. We shall here take a broad overview of these consequences[112], remembering that GDP is the sum of values added in all productive activities taking place in the area covered, whereas production of a firm is normally measured by the value of its gross output (the difference between the

112 The various points briefly surveyed in this subsection are developed in S. Basu, Intermediate goods and business cycles: implications for productivity and welfare, *American Economic Review* (June 1995).

value of gross output and value added is equal to the value of material inputs consumed in production: energy, materials, manufactured parts embodied in the output or any other goods and services used in the production process). Let us begin with some stylized facts before considering their explanation.

It is an old idea that, in industrial economies, roundabout methods of production and high specialization of firms are responsible for larger output fluctuations than would otherwise be observed. It is also an old idea that price changes are relatively less important, compared to quantity changes, in complex manufactured goods. These two ideas, which motivate the attention to be given to intermediate production, are not directly tested in the econometric work to be now reported, but they are consistent with its results.

A number of stylized facts were obtained by S. Basu (op. cit.) who made a statistical analysis of a panel of annual observations on 21 US manufacturing industries from 1959 to 1984. Quite significant correlations were found between the rates of change of different variables from one year to the next. The first such correlation concerns the input–output ratio for the aggregate of all material inputs. Let M_{it} be the aggregate intermediate input in industry i and year t. The input–output ratio then is M_{it}/y_{it} where y_{it} is gross output. When a linear regression with a constant specific to the industry is fitted for the rate of change of this ratio on the rate of change of y_t, the aggregate gross output of manufacturing, a value of 0.2 appears for the regression coefficient: the consumption of material inputs per unit of output is relatively more important in booms than in recessions. This is confirmed by a similar regression computed from the values of the ratio M_{it}/L_{it}, L_{it} being labour input. The regression then involves several regressors; the first one, the rate of change of y_{it}, appears with the coefficient 0.4; the additional regressors are introduced in order to capture the effect of labour hoarding and so to make this coefficient meaningful; they concern: the ratio of production to non-production workers, the average number of overtime hours worked and the number of hours worked by an average worker in each industry; these additional regressors have indeed positive coefficients.

A similar regression was computed concerning the ratio of the price w_{it} of the labour input to the price $p_{m,it}$ of the material input; the regressor then being the rate of change of y_{it}, a coefficient of 0.2 was again found. This shows that the relative price of intermediate inputs with respect to labour is countercyclical: high in recessions, low in booms. In order to avoid misinterpretations, let us note, moreover, that presenting these results their author adds that prices of intermediate goods are procyclical, but less so than labour costs.

Knowing these stylized facts and a few others, Basu develops an argument taking the imperfection of competition into account and showing the role of intermediate goods in the strength of the productivity cycle. Let us consider the main points in the argument.

In the first place, since competition is not perfect, prices exceed marginal costs and the equilibrium is inefficient. But prices are not fully flexible because of menu costs affecting their changes (and because of the fact that, in a neighbourhood of the profit-maximizing value of the output price, a small revision of this price implies just a second-order small change in profit). Rigidity of the output price means that the excess of the price over the marginal cost (the markup) is smaller in booms than in recessions because the marginal cost is then higher, as we saw in Section 6.4 of Chapter 8. Hence, the equilibrium is less inefficient in booms than in recessions.

This first conclusion holds independently of the presence of intermediate goods. But their presence adds two important features: an increase in price-rigidity and a distortion in the input-mix of respectively labour services and intermediate inputs. The increase in price-rigidity is explained as follows by Basu: "A firm raises its price only to the extent that its profits are squeezed between a fixed output price and rising input costs. With intermediate goods in production, the increase in firms' costs depends on whether other firms raise prices;... each firm simply waits by the mailbox to see if other firms have raised their prices". Such a reasonable strategy leads to staggering price changes along interindustrial exchanges and finally to spread them out over longer periods than if no intermediate input was used. Price rigidity and its effects are so magnified.

Also, prices of intermediate inputs are too high with respect to labour costs since intermediate goods are sold in markets subject to monopolistic competition. This distortion leads to too small inputs of intermediate goods and to too large inputs of labour. Since the markup is smaller in booms than in depressions, the relative price of intermediate goods with respect to labour will also be smaller in booms, correlatively the excess use of labour will be smaller, finally labour productivity will be higher.

We end up with a new reason for the productivity cycle independently of what "labour hoarding" may already imply. We also note that the argument rationalizes the two stylized facts derived from Basu's econometric analysis: in booms the input–output ratio of intermediate goods is high and the ratio between their price and the labour cost is low.

3. During depressions many plants shut down, which would have survived under better overall business conditions. This is the main reason for the procyclicality of net business formation, which was noted in particular by the National Bureau of Economic Research (Zarnowitz, op. cit.). The shutdown of those plants means a *decrease of productive capacities*, but among those

that were the least productive, because their techniques of production were surpassed by more efficient ones, or because the demand for their output was served at a lower cost by foreign producers, or still because customers had found interesting to shift their demand towards new goods or services, better suited to their needs. This is why the destruction of productive capacities is often presented as a cleansing of the productive structures, so relieved from obsolete elements and ready for the creation of much better performing capacities. The phenomenon can be seen from the other side: the competition for new products, new technologies, new sources of supply, or new types of organization plays an importance role in economic growth; it means replacement of old structures by new ones, this is "creative destruction" according to the phrase invented by J. Schumpeter.

Within manufacturing the phenomenon is particularly clear and was recognized as a major factor of productivity growth. A study of a sample of 22 US industries indicates that, over the period 1972–1987, aggregate growth was made up of 3.5 per cent due to plant-level technical progress, but of 6.7 per cent due to "reallocation", i.e. to the displacement of labour and capital from some plants to others, particularly to new ones[113].

The phenomenon is related to what was observed in many countries on labour market flow data: job destructions are counter-cyclical and, where firing is not restricted, their fluctuations have larger amplitude than that of the pro-cyclical fluctuations in job creations[114]. It is interesting to note moreover that the asymmetry of the business cycle, between recessions and expansions, does not clearly appear in data on job creations, but is in contrast made more acute in data on job destructions.

The reasons for the form taken by the cleansing effect of recessions have been studied by R. Caballero and M. Hammour[115]. They worked out a simple model, which seems to be well suited for the purpose and will be here briefly examined.

Cleansing is the consequence of the fact that techniques of production and specificities of output are embodied in productive capacities, which may not only decay because of progressive or repeated deteriorations, but also become obsolete. In order to deal correctly with this fundamental feature, Caballero and

[113] M. Baily, C. Hulten and D. Campbell, Productivity dynamics in manufacturing plants, *Brookings Papers on Economic Activity* (Microeconomics, 1992), pp. 187–249.

[114] For international data about fluctuations in job creations and job destructions, and for the role of firing restrictions see P. Garibaldi, Job flow dynamics and firing restrictions, *European Economic Review* (February 1998).

[115] R. Caballero and M. Hammour, The cleansing effect of recessions, *American Economic Review* (December 1994).

Hammour accept the extreme hypothesis of fixed proportions. A production unit of vintage t_0, i.e. created at time t_0, embodies the then leading technology and produces the same constant flow $A(t_0)$ of output, at the constant current cost 1, throughout its lifetime. Technical progress makes the productivity $A(t)$ grow at the constant exogenous rate $\gamma > 0$. Time t being continuous in the model, we write:

$$A(t) = A(0)\,\mathrm{e}^{\gamma t}. \tag{126}$$

Since production units that were created at different times (and thus have different productivities) coexist, we must keep track of their age distribution. Let $f(a,t)\,\mathrm{d}a$ be the number of units existing at time t which have age between a and $a + \mathrm{d}a$, for an infinitesimal $\mathrm{d}a$ (all functions are assumed continuous and differentiable in the model). Investment $I(t)\,\mathrm{d}t$, that is the number of units created between t and $t + \mathrm{d}t$, is equal to $f(0,t)\,\mathrm{d}t$ (the rate at which new units are created then varies from $f(0,t)$ to $f(0,t+\mathrm{d}t)$; but terms in $(\mathrm{d}t)^2$ must be neglected). If $\bar{a}(t)$ is the maximum age for units still existing at time t, then $f(a,t) = 0$ for any $a > \bar{a}(t)$. This maximum age is determined by the exogenous process of obsolescence, which we shall study in a moment. Production units are also subject to an exogenous decay at a constant rate δ. This means:

$$f(a + \mathrm{d}t, t + \mathrm{d}t) = \mathrm{e}^{-\delta \cdot \mathrm{d}t} f(a,t) \tag{127}$$

if $a < \bar{a}(t) - \mathrm{d}t$ (assumptions are made so as to rule out obsolescence between t and $t + \mathrm{d}t$ of units that have age smaller than $\bar{a}(t) - \mathrm{d}t$, see below). For each production unit, future decay will occur at random according to the Poisson process with probability δ.

In order to understand the economics of obsolescence we must introduce the positive price $p(t)$ of output at time t and consider the flow of return that is earned on units which were created at time $t - a$ and have not decayed before t. This flow is equal to $p(t)A(t-a) - 1$. Given (126) it is a decreasing function of age a reached at time t. We moreover assume that:

$$p'(t) < 0 \tag{128}$$

and conclude that, at time t, $\bar{a}(t)$ is determined by:

$$p(t)A\big[t - \bar{a}(t)\big] = 1. \tag{129}$$

Indeed, given (128), units of vintage $t - \bar{a}(t)$ would never in the future gain positive earnings and, given (126) and the differentiability of $p(t)$, units of younger vintages gave positive earnings in the past and will still give a positive earning for some time (the risk of decay during the next infinitesimal future period is infinitesimal, as well as is the decrease in earnings).

The price function $p(t)$ is obviously an endogenous element of the model. How is it determined? It must be such that the cost of investment, about which we were silent thus far, is covered by the discounted value of expected future earnings. The model assumes that this discounted value just covers the initial cost. The assumption will follow, for instance, from the hypothesis of free entry into, and exit from, the industry.

The earning which, at time t_0, is expected to be gained at time t from a production unit of vintage t_0 is nil if $t - t_0 > \bar{a}(t)$. Otherwise it is:

$$\pi(t_0, t) = \mathrm{e}^{-\delta(t-t_0)}\big[p(t)A(t_0) - 1\big]. \tag{130}$$

The multiplier of the square bracket is indeed the probability that the unit will not have decayed before or at time t. Let c be the investment cost of the production unit. Assuming for simplicity that the discount rate r is constant, the investment cost is exactly covered when:

$$\int_{t_0}^{t_0+\bar{a}} \mathrm{e}^{-r(t-t_0)}\pi(t_0, t)\,\mathrm{d}t = c, \tag{131}$$

where it is assumed that the owner of the production unit exactly forecasts the age $\bar{a}$ at which its production unit will have to be scrapped for obsolescence if it has not decayed before.

The right-hand member of (131) is written here as if the investment unit cost was not only constant, whatever t_0, but also fixed independently of how many units $I(t_0) = f(0, t_0)$ would be simultaneously invested. As we shall see in a moment, the assumption is not innocuous for the representation of the cleansing phenomenon. Worse, the assumption is not appropriate. However, it is an important contribution of the model to make us understand why the assumption is inappropriate. Thus, let us provisionally accept it.

In that case, there exist two constants $\bar{a}$ and $\bar{p}$ such that (129) and (131) are fulfilled with:

$$\bar{a}(t) = \bar{a}, \qquad p(t) = \bar{p}\,\mathrm{e}^{-\gamma t}. \tag{132}$$

Given (126), Equation (129) just requires:

$$\bar{p}A(0) = \mathrm{e}^{\gamma\bar{a}}. \tag{133}$$

Let us pose $t - t_0 = \tau$. Then (130) writes:

$$\pi(t_0, t) = \mathrm{e}^{\gamma\bar{a}-(\gamma+\delta)\tau} - \mathrm{e}^{-\delta\tau}. \tag{134}$$

Thus (131) determines $\bar{a}$ as the solution of:

$$\frac{1}{r+\gamma+\delta}\left[\mathrm{e}^{\gamma\bar{a}} - \mathrm{e}^{-(r+\delta)\bar{a}}\right] - \frac{1}{r+\delta}\left[1 - \mathrm{e}^{-(r+\delta)\bar{a}}\right] = c. \tag{135}$$

Looking for comparative assessments about this solution, the reader can check that both $\bar{a}$ and $\bar{p}$ increase if the investment cost c increases, and that $\bar{a}$ decreases (obsolescence comes earlier) if γ increases (technical progress is stronger). Other comparative results seem to be ambiguous.

Let us now reflect on the meaning of the outcome of this model, for the case of a constant and fixed unit investment cost. We found a solution in which the price of output decreases at the pace at which productivity increases (this is fine), and in which the scrapping age is determined independently of what business conditions can be, independently of whether demand is booming or not when scrapping occurs. In other words scrapping at time t reflects only the importance of investment at time $t - \bar{a}$. A boom of demand simply requires high new investment, not only for replacing decay and scrapping of old capital but also for building new capacities. Job creations have then to be strongly procyclical. Job destructions are insulated from current business conditions. All that is precisely the opposite of the observations from which we started. We were looking for an explanation of a phenomenon that our model seems to rule out.

But the observed phenomenon can be understood when we reflect about the failure of our first attempt. Indeed, an economy in which product and process innovations constitute the main reason for investing is also an economy in which the times for investment are not much flexible. Investments occur most often when projects are ready. We must find a way to incorporate in our formalization this lack of flexibility. Clearly, the easiest way for doing so is to assume that the unit investment cost c is an increasing function $c[I(t_0)]$ of the volume of investment made at time t_0.

This change in Equation (131) makes the model more satisfactory because we intuitively understand that the increase in $c(I)$ will smooth the creation process $I(t)$, which we found to be too procyclical with the hypothesis of a fixed cost. Correlatively the destruction process will be made procyclical. The model will be made more satisfactory but also much more complex: $I(t_0)$ is not exogenous and has to be explained, notably in relation to the state of demand at time t_0 and to the cost implications of reactions to this state of demand. There is a trade-off between the timing of creations of new capacities and the timing of destructions of old ones. Presence of this trade-off affects the whole evolution of these capacities: it affects it in the model as well as in fact.

Studying their model so revised, Caballero and Hammour find that it can be reduced to a system of two non-linear differential equations in $I(t)$ and $\bar{a}(t)$, which are also differential equation incorporating time-varying delays. The solution is analytically intractable. It can only be solved numerically by a process of successive approximations. Assuming a linear function $c(I) = c_0 + c_1 I$, and a continuous fluctuating process generating the demand function addressed to the representative firm, the authors give the results of simulations. These simulations concern first the case of sinusoidal fluctuations in demand, second a case in which fluctuations have the kind of asymmetry found in actual business cycles (shorter recessions than expansions).

Only a small elasticity of the creation cost $c(I)$, around 0.2, is needed to explain a stronger counter-cyclicality of destructions than the procyclicality of creations, as is observed in countries where firing is not restricted. Asymmetry is smoothed in the creation process, magnified in the destruction process (the model incorporates the idea that creations depend on expectations of future demand over the expected service-time of new equipments, whereas the timing of destructions depends mainly on current conditions). Given their specification, which assumes in particular a unit price-elasticity of the demand function, the authors find that creations $I(t)$ declines only mildly in response to sharp contractions and that the price $p(t)$ falls more sharply, a reaction which magnifies destructions.

2.9. Unemployment persistence

In a macroeconomic textbook, such as this one, unemployment may not receive the complete treatment it deserves. It is not normally viewed as a feature of long-run growth. Thus, it should come in the chapter dealing with business fluctuations. Variations in unemployment are indeed part of the business cycle and may be studied as such. But they often are more than that. In some cases the impacts of booms and recessions seem to be superimposed on a movement with longer duration. This is particularly so now in Western Europe, which offers in this respect a striking contrast with the US.

If the image is correct, we should like to have a good understanding of such a longer-run movement, that is, of the set of phenomena responsible for it. Alternatively, we should like to understand why, particularly with respect to unemployment, the dynamics over two decades or so may differ from the image of a simple business cycle.

The concern will be present in this section which will adopt a broad conception of unemployment persistence. A narrow conception would see persistence as a characteristic of the unemployment stochastic process taken in isolation. An extremely simple model of this process would state:

$$u_t = \beta u_{t-1} + \gamma + \eta_t, \tag{136}$$

where β and γ would be two numbers and the unobserved η_t would be a white noise process. The value of β would then provide a natural measure of the degree of *persistence*. In the case $\beta = 1$ we could even speak of *hysteresis*, since the mark of a transitory shock η_t would remain unchanged in all subsequent unemployment rate $u_{t+\tau}$ (for $\tau > 0$). We shall not try to spell out what should more generally be the definitions of persistence and hysteresis. Present usage of these words remains too uncertain for making an attempt at codification useful.

We shall give some account of research dealing with the lag of the unemployment rate in its adaptations to shocks that are maintained unchanged for several periods, a lag that was recognized in Chapter 8 when we dealt with the short-run dynamics. But we should like to do more and to understand whether and how persistence may become a more serious problem after a series of adverse shocks.

The question was well posed by O. Blanchard and L. Summers in their examination of the European unemployment problem[116]. They wrote in their introduction: "The sustained upturn in European unemployment challenges the premise of most macroeconomic theories that there exists some "natural" or "non-accelerating inflation" rate of unemployment towards which the economy tends to gravitate... The European experience compels consideration of alternative theories of "hysteresis" which contemplate the possibility that increases in unemployment have a direct impact on the "natural" rate of unemployment" (p. 15). And towards the end of their study they still recognize the difficulty of the challenge: "Our model of persistent unemployment may explain important aspects of the current European depression and the very different behavior of European and American labor markets. The evidence presented so far leaves

[116] O. Blanchard and L. Summers, Hysteresis and the European unemployment problem, in: S. Fisher (ed.), *NBER Macroeconomics Annuals 1986* (MIT Press, 1986).

open a crucial question, however. Is the presence of hysteresis in European unemployment a consequence of the heavily regulated and unionized character of European labor markets? Alternatively, is hysteresis the result of a sequence of adverse shocks to employment? The case that major structural reforms are needed if full employment in Europe is to be restored depends on an affirmative answer to the first question, while the case for expansionary macroeconomic policies is more compelling if the second question can be given a positive answer" (pp. 61–62).

Readers of this book should be better prepared than others to the idea that macroeconomic theories may exist according to which the rate of unemployment does not need to gravitate towards a natural rate. Before we started discussing the concept of such a gravitation in Chapter 8 we indeed had already met the theoretical possibility of alternative dynamic behaviours in Chapter 7, particularly in Sections 5.4 to 5.6. We may also remember at this point that, in Part 6 of the same Chapter 7, we surveyed a number of theories meant to explain why unemployment might tend to systematically exceed frictional unemployment. Some of these theories also served in attempts at explaining unemployment persistence.

In this section we shall insist both on empirical evidence and on the complexity of the persistence phenomenon. We shall not rule out theories, but try to put them in context. Discussing at the end the causes of the European unemployment problem will be natural, given such a purpose.

2.9.1. *Empirical macroeconomic descriptions*

To begin with a look at relevant data, the contrast between Western Europe and the US is exhibited on Figure 7 where are plotted the two series of the unemployment rates, Western Europe being defined as the group of the fifteen countries now making the European Union. The figure is so well known and so clear that it needs no other comment than reminding that, starting in 1960 and up to the end of the 1970s, at least, it was common to hear it said in Europe that the US economy was unable to solve its unemployment problem, each peak of its unemployment rate being since 1950 higher than the preceding ones.

We must also realize the variety of experiences within Western Europe. For instance, all through the 1960s, the Irish unemployment rate was highest in Europe; it was surpassed by the Spanish rate in 1979, but still reached 17 per cent at its peak in 1985; at 6 per cent in 1999, the Irish rate is now surpassed in nine countries of the European Union. The Swedish unemployment rate never reached 4 per cent throughout the 1960s, 1970s and 1980s, but jumped from 2 to 10 per cent between 1990 and 1997. Since 1984 the difference between the Spanish rate and the Portuguese rate amounts to 12 points or more.

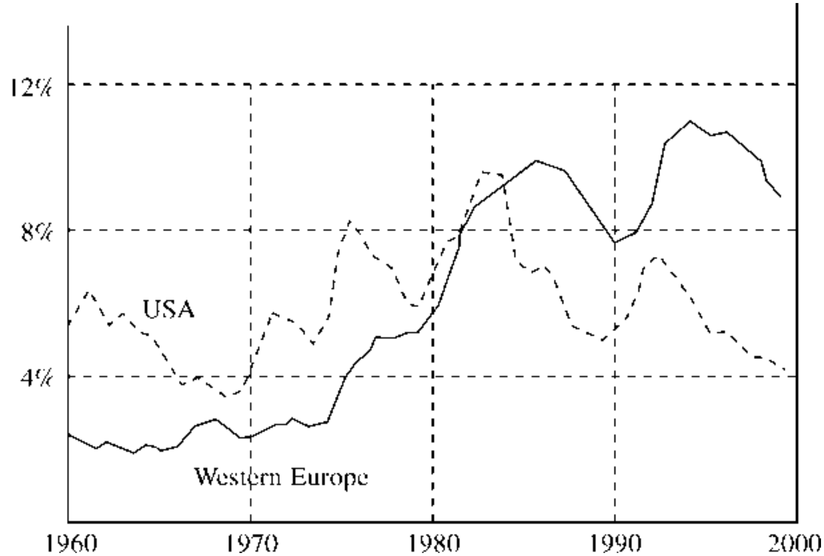

Fig. 7. Unemployment rates.

Asymmetry between upward and downward phases is noticeable in Figure 7, the first phases being typically shorter than the second. This can be seen as a mirror image of the asymmetry of the output series, where contractions are shorter than expansions, as we saw in Section 2.7. However, a stronger asymmetry would be implied by the idea of a persistence that would be specific to unemployment and would follow from an asymmetric ratchet effect, each increase in unemployment engaging a new level in the resistance to labour-market clearing.

Empirical studies, with which we begin, bear either directly on series of the aggregate unemployment rate or, as we shall see in a moment, on a more disaggregated examination of the proximate causes of variations in unemployment.

When wondering about the perplexing persistence of European unemployment, analysts often thought it useful to attribute observed changes in unemployment rates to *changes in "the natural rates" and/or to changes in "the cyclical rates"*. According to A. Jaeger and M. Parkinson[117], two main classes of explanatory hypotheses could be using the distinction. The first class would

[117] A. Jaeger and M. Parkinson, Some evidence on hysteresis in unemployment rates, *European Economic Review* (February 1994).

assume that the natural and cyclical rates evolve independently, the natural rate being taken as dependent on "structural factors" (excess real wages, mismatch between the compositions of labour supply and demand, unemployment benefit system, ...). The second class would embody an idea of hysteresis: cyclical movements in unemployment would somehow be propagated to the natural rate (because of erosion of the human capital of the long-term unemployed, because of selfish behaviour of insiders, ...). In order to assess the relative importance of the two classes of hypotheses, Jaeger and Parkinson define an unobserved-components econometric model of the unemployment rate. Let us look at the model and at their findings.

Let u_t be the observed unemployment rate and its two unobserved components be u_{nt} for the natural rate, u_{ct} for the cyclical rate:

$$u_t = u_{nt} + u_{ct}. \tag{137}$$

Underlying the distinction is the idea that the cyclical rate may be taken in this empirical approach as ruled by a stationary stochastic process, which is assumed to be second-order autoregressive (the authors report that attempts with higher order processes did not significantly improve the fit):

$$u_{ct} = \varphi_1 u_{c,t-1} + \varphi_2 u_{c,t-2} + \varepsilon_{ct}, \tag{138}$$

where ε_{ct} is ruled by a white-noise Gaussian process with variance σ_c^2. As for the natural rate it would follow a random walk if it was not for the propagation coming from the cyclical rate. So, the authors assume:

$$u_{nt} = u_{n,t-1} + \alpha u_{c,t-1} + \varepsilon_{nt}, \tag{139}$$

where ε_{nt} is ruled by a white-noise Gaussian process, independent of $\{\varepsilon_{ct}\}$ and with variance σ_n^2. The estimated value of α will be taken as a measure of the strength of the propagation from the cyclical to the natural rate.

The model defined by (137) to (139) is very simple. It does not claim to provide an explanation of why the unemployment rate varied, but rather to rationalize the empirical breakdown of u_t into two meaningful components. We note that the model does not recognize any asymmetry in movements up or down of unemployment. Persistence appears, in the extreme form of hysteresis, by the random walk part of the natural rate and, again but in a milder form, by the sum of the two coefficients φ_1 and φ_2 in the stochastic process of the cyclical rate.

A trouble with the simple specification (137) to (139) comes from the fact that it does not permit identification[118] of φ_1, φ_2, α, σ_c and σ_n from simply the observation of the series of the unemployment rate u_t. Information given by an additional observed variable, which is known to be related either to the cyclical or the natural rate component, is needed. But can any additional observed variable be introduced in specification (137) to (139) without making the model essentially more complex?

The authors argue that the mean-adjusted real growth rate of the aggregate output of the G-7 countries provides the required information. Indeed, they say that the value of this growth rate can be taken as independent of the natural rate of anyone country, because this natural rate was determined by factors specific to the country[119]. In order to introduce the new variable (g_t, say) into their specification the authors assume:

$$g_t = \beta g_{t-1} + \delta\left(u_t^c - u_{t-1}^c\right) + v_t, \tag{140}$$

where v_t is meant to be ruled by a white-noise Gaussian process, independent of $\{\varepsilon_{ct}\}$ and $\{\varepsilon_{nt}\}$, with variance σ_v^2.

Taken literally the model, which was used for estimation on the basis of Equations (137) to (140) and related assumptions listed above, is bizarre. It is recursive and means that, in the application to each country, the G-7 variable g_t is determined from the evolution of its national cyclical unemployment rate. But we must recognize that a difficulty is common to all empirical approaches to the study of macroeconomic phenomena: for one reason or another their underlying models are all exposed to the criticism of being misspecified; fitted results are, however, interesting to know. In this spirit let us look at results obtained by Jaeger and Parkinson.

These authors applied their empirical approach on seasonally-adjusted quarterly data, covering the years 1961 to 1991, and this for four countries: Canada, Germany, the UK and the US. The stationary stochastic processes of the cyclical components were found to be very similar, with φ_1 close to 1.6 and φ_2 close to -0.7, implying fluctuations close to a damped cycle with a period of 6 years. Neither the similarity nor the nature of the processes is surprising considering in particular the identifying role of the common variable: the G-7 output growth rate.

[118] The econometric concept of identification was already met in Section 5.4 of Chapter 8 about the VAR methodology. It will be met again in the next part of the chapter about simultaneous-equation models. The case of unobserved-component models is a third example.

[119] The authors report that, in an earlier version of their article, they had used the national capacity utilization ratio as the identifying additional variable and that the estimation results were "not too much" different.

For the US the coefficient α measuring the force of the propagation from the cyclical to the natural rate, what the authors call “the hysteresis parameter”, does not differ significantly from zero. Simultaneously the estimated cyclical component closely tracks the deviation of the unemployment rate from its average value over the sample. It is probably a reflection of two facts: the unemployment rate exhibits hardly any visible long-run trend, the US has a comparatively high weight in the G-7 output growth rate.

In contrast, the coefficient α is positive and quite significant for the other three countries, with the result that the smoothed “natural rate” follows quite closely the actual unemployment rate. From a data analysis viewpoint, the result can be explained: whatever the smoothing, the evolution followed by the unemployment rate during the three decades, particularly in the two European countries, was very far from what simulations of a random walk would be expected to give. But from a macroeconomic viewpoint, how should we interpret the finding? It would be, I fear, rash to take it as evidence of the force of the various phenomena that have been associated with the idea of unemployment hysteresis and have been embodied in specific theories of unemployment equilibrium. Should not we rather keep looking for causes which, since the middle of the seventies, have strongly disturbed the macroeconomic equilibrium through channels that have since long been taken as responsible from business fluctuations?

Another application of a purely empirical approach is due to M. Bianchi and G. Zoega[120]. It differs in two respects: the data base and the model. The data are annual, but bear on a much longer period, a century or more instead of three decades[121]. They concern three countries: the UK (1855–1994), the US (1890–1994) and France (1894–1994).

Bianchi and Zoega consider that *the natural rate of unemployment may significantly shift on occasions from period to period*, for reasons which they nei-

[120] M. Bianchi, G. Zoega, Challenges facing natural rate theory, *European Economic Review* **41** (1997) 535–547.

[121] Long historical macroeconomic series are notoriously subject to lapses from intertemporal and/or international comparability. This particularly applies to unemployment series, because of the sizable impact of small changes in the definition of the concept or in the method used for the collection of elementary data. In the present case, I consider the French series as unreliable for comparison of the recent period with the first half of the century where the series is grossly underestimated (the reasons are well explained in R. Salais, N. Baverez and B. Reynaud, *L’invention du chômage* (Presses Universitaires de France, Paris 1986)). Similarly the comparative levels of the rates in the UK and France appear suspect before 1950. Whereas British unemployment was known to have been high in the 1920s, the historical evidence does not support the idea of a much lower unemployment in France than in England during the 1930s or the first period of the century. However, these lapses from comparability are unlikely to have much affected the results to which the authors give attention.

ther try to explain nor relate to changes in the observed rate. In our notation we then have a natural rate $u_n(s_t)$ function of a latent state variable s_t, which is assumed to follow a Markov stochastic process: with given probabilities, s_{t+1} is equal either to s_t or to anyone of a number of other possible values. The authors then simply pose:

$$u_t = u_n(s_t) + \beta u_{t-1} + \eta_t, \tag{141}$$

where the parameter β characterizes the degree of unemployment persistence and the unobserved error η_t follows a Gaussian white noise process.

In order to estimate β, the dates and natures of the shifts of s_t and the values $u_n(s_t)$, the authors applied the same method as was used by J. Hamilton for the work presented in Subsection 2.1.8 of this chapter (op. cit.). In most cases the shifts of s_t were distant by 10 years or more, so that periods during which the natural rate could be assumed roughly constant were identified. Except for a few shorter spans the following periods were identified: for the UK, 1855–1870, 1876–1895, 1901–1914, 1919–1940, 1941–1974, 1975–1994; for the US, 1890–1898, 1899–1907, 1914–1925, 1930–1942, 1943–1994; for France, 1895–1930, 1931–1942, 1943–1967, 1968–1980, 1981–1994. It is interesting to compare these determinations of periods during which the natural rate could be assumed constant with the estimations of the natural rate by Jeager and Parkinson: the US rate was indeed then found about constant from 1961 to 1991 (the full sample period); the UK rate was found to have been close to 3 per cent from 1961 to 1974 and to have traced subsequently (up to 1991) a large wave around 7 per cent. We may say that, for the countries and periods which were covered by both studies, the agreement is pretty good. But the longer historical perspective suggests that approximate constancy of the US natural rate of unemployment, as detected by an empirical analysis of the observed rate, is a special feature of the postwar.

The estimate of β found by Bianchi and Zoega amounts to 0.4 for the US and to 0.6 for the two European countries. The match with the study of Jaeger and Parkinson is less good. First, no cyclical behaviour of the unemployment rate appears in the fit, except for the effects of shifts from one value of the state variable to another (which seems to concern more “long cycles” than the more commonly discussed business cycles). Second, the results assert that unemployment persistence is stronger (β significantly higher) in the two European countries than in the United States, whereas Jaeger and Parkinson found similar values of $\varphi_1 + \varphi_2$ in the four countries they studied (the excess over the US were 0.009 for the UK and 0.016 for Germany, both probably not significant).

Overall, this author is tempted to state that the two empirical macroeconomic descriptions are complementary.

2.9.2. *Labour force heterogeneity*

Moving towards more explanatory studies of unemployment persistence we have to analyse labour market flows, which were already introduced in Section 8.4 of Chapter 8. We have also to realize the importance of labour force heterogeneity, an importance which may be overlooked when analysis bears directly on wholly aggregated data. For this purpose we shall refer to an article by M. Darby, J. Haltiwanger and M. Plant[122], which is also interesting in other respects for us here. The article is focused on three proximate groups of determinants of unemployment dynamics: the rates at which workers enter into unemployment, the rates at which unemployed people exit from unemployment and the rates at which numbers of people belonging to the labour force grow. For an homogeneous group of people, which are all alike except with respect to their employment status, the relation between these concepts and changes in unemployment is simple. It becomes more complex when heterogeneity must be taken into account. So, let us begin with the hypothetical case of identical individuals.

Let N_t, U_t, EU_t, OU_t be the numbers of people who in period t respectively belong to the labour force, are unemployed, enter into unemployment, go out of unemployment. Let us consider the following ratios[123]:

$$\begin{aligned} u_t &= \frac{U_t}{N_t}, \qquad \varphi_t = \frac{EU_t}{N_t}, \\ \pi_t &= \frac{OU_t}{U_{t-1}}, \qquad \gamma_t = \frac{N_t - N_{t-1}}{U_{t-1}}. \end{aligned} \tag{142}$$

We call then respectively the rate of unemployment, the rate of entry into unemployment, the rate of exit from unemployment, the rate of growth of the labour force. In this case of identical individuals, we may interpret the two first ratios as probabilities applying indistinctly to any member of the labour force; we may similarly interpret the third ratio as the probability for any unemployed individual to find a job or to leave the labour force.

[122] M. Darby, J. Haltiwanger and M. Plant, Unemployment rate dynamics and persistent unemployment under rational expectations, *American Economic Review* (September 1985).

[123] We use here for convenience a notation close to the one adopted in the article discussed. The ratio φ_t and π_t were denoted respectively as e and p in Section 8.4 of Chapter 8.

Simple algebra leads to the following identities:

$$U_t = U_{t-1} + EU_t - OU_t$$
$$u_t - u_{t-1} = \varphi_t - \frac{\pi_t + \gamma_t}{1 + \gamma_t} u_{t-1}. \tag{143}$$

It is convenient to introduce the "growth-adjusted probability of leaving unemployment":

$$\psi_t = \frac{\pi_t + \gamma_t}{1 + \gamma_t} \tag{144}$$

and to write (143) as:

$$u_t - u_{t-1} = \varphi_t - \psi_t u_{t-1}. \tag{145}$$

Clearly, if the ratios φ_t, π_t, γ_t remain constant and equal to $\bar{\varphi}$, $\bar{\pi}$, $\bar{\gamma}$, then ψ_t also remains constant at $\bar{\psi} = (\bar{\pi} + \bar{\gamma})/(1 + \bar{\gamma})$ and the unemployment rate u_t is similarly constant at $\bar{u} = \bar{\varphi}/\bar{\psi}$. It is then natural to consider the deviations from a consistent set of values $\bar{\varphi}$, $\bar{\pi}$, $\bar{\gamma}$, $\bar{\psi}$ and $\bar{u}$ taken as representative of what is found along long-run trends. Hence we pose:

$$\hat{\varphi}_t = \varphi_t - \bar{\varphi}, \qquad \hat{\psi}_t = \psi_t - \bar{\psi}, \qquad \hat{u}_t = u_t - \bar{u}. \tag{146}$$

Finally, in order to exhibit how shocks on φ or ψ affect the dynamics of the unemployment rate, Equation (145) is rewritten as:

$$\hat{u}_t = \hat{\varphi}_t - (\hat{u}_{t-1} + \bar{u})\hat{\psi}_t + (1 - \bar{\psi})\hat{u}_{t-1}. \tag{147}$$

This last equation may be seen as ruling the unemployment rate dynamics in the case of full homogeneity. We indeed read in (147) how, for instance, a temporary shock in the probability of entering into unemployment, i.e. a shock in $\hat{\varphi}_t$ just during period t, affects the unemployment rate when ψ_t remains fixed at the long-run value $\bar{\psi}$. In period t there is an immediate impact by the same amount as the shock, but the effect persists beyond the period because of the last term in (147). However, the effect decays fairly quickly. According to Darby et al., only 15 per cent of the shock would still be felt in the unemployment rate after three more months. This figure concerns the US for which the rate of exit from unemployment is notoriously high, much higher than in most other countries, mainly because of institutional differences in the functioning of the labour market. A lower value of $\bar{\psi}$ implies more persistence, as is evident on Equation (147). But nowhere would unemployment persistence be a serious problem if this equation, applied directly at the aggregate level, provided an accurate account of unemployment dynamics.

The three authors of the article under discussion indeed show that heterogeneity makes persistence more serious. But before we turn to their argument we may question whether there are not other reasons for more persistence than shown in (147). This equation looks like an accounting relation implied by definitions of the concept, hence a relation totally immune from the interference of behaviours or adjustment costs. It is indeed an accounting relation, but one to which we give an interpretation that is not so immune. Indeed we note that, in π_t as defined, there is a lag of the denominator with respect to the numerator, and we interpret π_t as the relevant probability to consider when dealing with exit from unemployment. This is tantamount to assuming that there is a (small) lag between the decision that will withdraw persons from unemployment and the moment at which they will no longer be recorded as unemployed. This is not unrealistic. But should not other lags be recognized as realistic and as explained by behaviours or adjustment costs? We leave the question unanswered at this stage. But its relevance will appear later in this section when we shall discuss the full range of reasons which have been given in order to explain persistence.

The three authors want to convey a quick understanding of the effect of heterogeneity. For so doing they assume that workers are divided into two groups. "Those in the first group have little firm-specific human capital and they experience unemployment frequently, but the length of these spells are brief... Members of the second group rarely experience unemployment, but when it occurs, search is extensive and well supported by unemployment compensation, other family income, and assets". For the first group the relevant probabilities φ_1 and ψ_1 are large; for the second φ_2 and ψ_2 are both low. The authors work out the equations applying to this case and show that the long search of the second group will lead to a much larger persistence than that based on Equation (147) in which average coefficients would be used, ψ_t in particular being defined as:

$$\psi_t = \frac{N_{1t}\bar{u}_1}{N_t\bar{u}}\psi_{1t} + \frac{N_{2t}\bar{u}_2}{N_t\bar{u}}\psi_{2t}. \tag{148}$$

After a recession causing mass disemployment, the low-probability group will be over-represented for a considerable period of time. During the recovery period, u_{1t} will quickly return to $\bar{u}$ and thereafter u_t will decrease about at the speed of u_{2t} multiplied by the share that unemployment of the second group will have in total unemployment.

Heterogeneity is responsible not only for an underestimation of persistence as commonly measured but also for errors leading to underestimate fluctuation in the cyclical component of the unemployment rate; those errors lead correlatively, in a decomposition such as given by (137), to find too much fluctuation in the natural rate of unemployment. The authors show how it is possible to estimate long-run values $\bar{\varphi}_i$, $\bar{\pi}_i$, $\bar{\gamma}_i$, $\bar{u}_i$ for various subgroups into which the labour force is broken down. They also show how to then derive slowly evolving aggregate normal values $\bar{\varphi}_t$, $\bar{\pi}_t$, $\bar{\gamma}_t$, $\bar{u}_t$ which are adjusted for heterogeneity bias (the natural rate of unemployment is then identified with $\bar{u}_t$). Correspondingly, the adjusted cyclical component $\hat{\varphi}_t$, $\hat{\pi}_t$, $\hat{\gamma}_t$, $\hat{u}_t$ may be obtained.

In the article the method is applied to a breakdown of the US labour force according to its age-sex composition. The authors mention also, as potentially relevant but not introduced in their calculation at that stage, the industrial or occupational composition of the labour force, and still more a distinction between two groups capturing the difference sketched in the example quoted above here ("permanent" or temporary attachment to jobs). Differences between the long-run computed characteristics are, however, striking, particularly for the probabilities $\bar{\varphi}_i$ of entry into unemployment during a month: about 9 per cent for the two age-groups 16 to 19 (respectively males and females), 5 per cent for the two age-groups 20 to 24, a minimum at 1.0 per cent for males aged 44 to 64 and at 1.3 per cent for females aged 55 and over. The adjusted normal values, hence in particular "the natural rate of unemployment", show little variation. Only for $\bar{\varphi}_t$ is an evolution worth mentioning: the probability slowly increase from 2.2 per cent in the late 1950s to 2.8 per cent in the late 1970s, then decreasing to 2.7 at the end of the series in 1982.

In its last section the article complement the descriptive measures so found with equations that explain the movements of π_t, φ_t and γ_t around their normal levels. Several messages are there issued by the authors: (i) heterogeneity is an important source of persistence, (ii) data appear to be consistent with equilibrium models of persistence because no evidence is found for the role of expectational errors, which should result in autocorrelated errors in the equations, (iii) in the simple model that was chosen for an attempt at explaining monthly changes in cyclical components, only two candidate explanatory variables were introduced, inventories and money (current and eleven lagged values of the level of inventories, eleven lagged values of the deviation of the money stock from its smoothed evolution); it then appeared that money shocks were not directly causing changes in cyclical unemployment.

More precisely, in the two equations meant to explain $\hat{\pi}_t$ and $\hat{\varphi}_t$, the money regressors were found to have an insignificant role, whereas the inventory regressors played a quite significant part with the expected signs. Additional variables meant to capture the effect of a changing heterogeneity were also found significant: in the equation meant to explain exit out of unemployment, it was the lagged proportion among unemployed of those whose current unemployment spell was shorter than five weeks (a high proportion meant a high probability of leaving unemployment, everything else being equal); in the equation meant to explain entry into unemployment, seven additional regressors were introduced, each one measuring the share, in the labour force, of workers of a specific industry (mining, . . . , finance and services): working in mining was found to imply a particularly high risk of unemployment and working in transports a particularly low risk.

2.9.3. *Heterogeneity of shocks and common features*

In reflections about analysis of the unemployment problem by applied macroeconomists, it is now customary to make a clear distinction between occurrence of shocks hitting the economy and their transmission to the main macroeconomic variables. Persistence is then viewed as a property of "the transmission mechanism". But the language may be the source of a confusion about the exact nature of the shocks, a confusion that may lead us to underestimate the true complexity of the persistence phenomenon.

This Part 2 of the present chapter shows how successful has recently become the recommendation made long ago by Ragnar Frisch, when he argued in favour of a stochastic approach to the study of business fluctuations. Most of our modern theories state the *existence of a stochastic shock process* whose properties may be specified in various ways, as we saw. But when we turn attention to a particular historical phase in order to explain what then happened, we often adopt a different conception of shocks, a conception which does not fit within that of unexpected shocks produced as the normal exogenous realization of a stochastic process with well specified properties. We rather argue that, at a given time in a more or less large geographical zone, *something occurred, which was exceptional and exogenous to the economic system*. Except for exogeneity the shocks which occurred at different times, even maybe in different zones, have little in common.

Let us pause, just in order to recall what happened in a few important cases. In 1973 and again in 1979–80 *the price of oil jumped up and simultaneously world aggregate demand was depressed* because important oil-producing countries were not in a position to quickly spend the large inflow of extra incomes they were receiving. In the late 1970s and early 1980s, after a long period of expansion turning into one of accelerating inflation, many firms in large European countries experienced *an exceptionally deep depression of their profitability* and had to cut the volume of their projects or even of their current operations. In the early 1990s, after an energic financial deregulation opening credit facilities to households and small firms and at the end of a speculative boom on the housing and financial markets, *many agents realized they had to deflate their debts* and, for so doing, to sell assets, to revise downward their investments or to cut on their current expenses. In the early 1990s, a political analysis led the German government to impose on the economy *an unexpected management of an unexpected unification*.

Speaking of shocks for such events is natural. But listing them suffices to realize that each one of them, except the second oil shock, was different from those which had been earlier experienced, so that it cannot be an accurate formalization to assume that they were all generated by a common stochastic process. For applied macroeconomists each shock was special: by its nature, its strength or its duration, and also by the location of its impacts on the economic system. There is little reason to believe that the same persistence properties would result in all cases from the transmission of the initial shock to the terminal rise in unemployment.

Against this historical background let us wonder about the relevance of the three econometric studies we discussed earlier in this section. The one by Darby, Haltiwanger and Plant is clearly addressed to the typical pattern of reactions to normal rather than to exceptional shocks. But it explains how labour force heterogeneity leads to longer lags than the consideration of individual cases might lead some to believe; longer lags mean higher values of the coefficient β in such simple representations as the ones given by Equations (136) or (141). The results are very likely to hold also after exceptional shocks, but only if we are able to know the impact of each such shock on the microeconomic variables φ_{it} and ψ_{it}. In any case those readers who are more interested in big things than in minute details may judge that characterizing unemployment persistence after exceptional shocks requires investigation of different experiences from that of the US during the 1960s and 1970s.

Reference to a shifting natural rate of unemployment in Subsection 2.9.1 opens possibilities for such investigations. But the simple descriptive schemes offered either by (137) to (139) or by (141) provide just some background aggregate evidence, which has little explanatory power. This is obvious with (141), since the Markov process ruling the transitions from state s_t to state s_{t+1} cannot claim to be more than a descriptive device; moreover, this process is not really estimated, considering the small number of transitions actually found and the potentially large number of modalities that $u_n(s_t)$ could take. With Equation (139) the model used by Jaeger and Parkinson contains the white noise process $\{\varepsilon_{nt}\}$ which appears suspect, considering what was exposed above, if we want to use it as a support for the identification of shifts in the natural rate.

Granted that essentially different types of shocks have to be recognized, it is still possible that, once important unemployment occurred, following whatever shock, *some features appear that are independent of the type of the initial shock and are themselves responsible for part of persistence* in subsequent unemployment. We may think here of two such features whose existence is hardly questionable but whose role on persistence is less obvious. These features concern respectively some individual workers, those remaining long or repeatedly unemployed, and the whole society in countries facing prolonged mass unemployment.

Being long unemployed leads to, often irreversible, *loss of part of one's "human capital"*, broadly defined. One loses the opportunity to maintain and update one's skill by working. One loses the habit of a high and regular level of activity. One loses confidence in one's abilities, up to the point that, in extreme cases, health is destroyed. On top of these direct effects on the person there are social disadvantages. In particular, having remained long unemployed is a serious handicap when applying for a vacant job; it is often a stigma in social relations. So the above effects have a cumulative character.

Anybody knows individual examples of that sort. But lacking results coming from truly scientific experiments, econometricians find it difficult to gauge the exact importance of the phenomenon. For instance, in order to measure the loss in human capital resulting from missing the opportunity of updating one's skill, we might think of referring to data showing how wage rates increase with seniority. But we would then have to allocate part of the increase to a deliberate policy of employers in the conception of the (often implicit) long-term contracts proposed to their employees: in order to sustain durable attachment of employees to their jobs employers usually offer careers in which remunerations increase with seniority for a number of years independently of any possible skill upgrading. Similarly, it was often argued that loss of experience and market stigma explain why the frequency of reentry into employment is all the lower as workers remained longer unemployed. But this may also be explained by heterogeneity of the pool of workers who lost their jobs: some look attractive for employers and will be soon reemployed; most of those remaining long unemployed had from the start unattractive attributes for employers. Moreover, given the available information, econometricians are hardly ever in a position to measure the importance of this selection bias in estimates of the loss of experience[124].

[124] See, for instance, J. Heckman, Identifying the hand of the past: distinguishing state dependence from heterogeneity, *American Economic Review* (May 1991).

Let us, however, accept the idea that long unemployment leads to a loss of human capital. Is the phenomenon a factor of unemployment persistence? This is indeed likely, but the factor is also unlikely to play an important part except where the unemployment benefit scheme gives indefinite entitlement to income compensation, no matter how long the worker may remain unemployed. Indeed, there is evidence of persistence in employment, coming from persistence in the labour supply, more precisely coming from the behaviour leading individuals to participate or not in the labour force. For example, the large increase in the participation of US adult women during World War II marked a permanent shift upward with respect to prewar trend[125]. Experience of long unemployment induces many persons to stop looking for jobs and thus to withdraw from the labour force, a behaviour which has the same kind of persistent effect as was observed, in the opposite direction, after World War II. But such withdrawals from the labour force are also, by definition, withdrawals from unemployment: they work against unemployment persistence as recorded.

However, where income compensation is granted to the unemployed for an indefinite period after the loss of employment, even though at a lower level of support, such withdrawals from the labour force occur much less frequently: indeed they are then associated with the loss of unemployment benefits[126]. But the real significance of this factor of unemployment persistence is questionable, since it mainly reflects how a given person reports his or her status when asked to do so, and this depending on his or her entitlement to income support[127].

Another common feature of long periods of unemployment, whatever their cause, is to generate pessimistic forecasts announcing that unemployment will never return to earlier levels. At least since the beginning of the 1990s public

[125] See on this point and more generally, K. Clark and L. Summers, Labour force participation: timing and persistence, *Review of Economic Studies* **XLIX** (1982) 825–844.

[126] Among the many econometric studies of the impact of unemployment insurance on unemployment, the most relevant here probably are: K. Clark and L. Summers, Unemployment insurance and labor market transitions, in: M. Baily (ed.), *Workers, Jobs and Inflation* (The Brookings Institutions, 1982) for results from microeconomic data; R. Layard, S. Nickell and R. Jackman, *Unemployment* (Oxford University Press, 1991), Section 9.5, for results from aggregate time series data.

[127] More significant is the fact, reported in Section 8.6 of Chapter 8, that the social norm of searching for a job appears less mandatory when and where being unemployed becomes more common.

opinions as well as intelligentsia in Europe firmly believe in such forecasts[128]. Blanchard and Summers (op. cit.) document a similar pessimism in the US during the Great Depression. They write in particular that "such pessimism was pervasive even among those charged with alleviating the situation" (p. 66). In such times most economists do not share this systematic pessimism; notably economic historians assert that unemployment indeed fluctuates, but with no long-run tendency to increase. But such assertions are not taken as credible.

Pessimistic expectations are favourable to the expansion of neither supply nor demand. As long as they are held, they depress economic growth. They should be recognized as a potentially strong factor of persistence. Notice, however, that this effect of depressing expectations is superseded if and when growth resumes: the rather widespread pessimistic expectations of the late 1940s and early 1950s did not prevent subsequent growth.

Before closing this subsection, let us add a comment. Persistence is a form of "history-dependence": what will be observed in five years from today will depend on what happened in the past up to now and on the future evolution up to that horizon. But there are various notions of history-dependence. In a weak sense any truly dynamic process would be history-dependent because each temporary equilibrium would depend on the values taken by some variables determined by previous temporary equilibria (extending somewhat the normal usage, we may speak of "stock variables" for referring to all such variables). However, the phrase "history-dependence" is most often meant to say more: a strong dependence would exist with respect to some past accident, even though the accident may have appeared small at the time. Typical of such a possibility are cases in which multiple temporary equilibria would exist for some combinations of values taken by stock variables, whereas the equilibrium would be otherwise uniquely determined and continuous as a function of the stock variables: a small accident in a neighbourhood of such a combination could then shift the subsequent evolution toward a different route from the one that would have been followed in the absence of the accident.

Some economists like to think that the evolution of unemployment may so depend on past shocks in a way that could not have been forecast, given the limitations of our knowledge. These economists belong to a subgroup of those stressing coordination failures (see the beginning of Section 6.3 of Chapter 7). Our short discussion of endogenous cycle in Section 2.6 above suggested how

[128] Among many other evidences, I may quote here the general tone running through the report of a group of experts working for the French Commissariat Général du Plan (J. Boisonnat (ed.), *Le travail dans vingt ans* (O. Jacob, Paris, 1995)). The main concern was to know how to adapt to a future in which there would not be enough paid jobs for everybody.

models exhibiting even unpredictable chaotic behaviour can be built. The author of this book has serious doubts as to the relevance of a vision that forbids diagnosis about the causes of unemployment persistence. He will go on studying here arguments that were given as explaining this persistence and the forms it can take.

2.9.4. *Wages in the transmission of shocks*

Let us now look at the literature on our subject, and wonder about which theoretical explanations are most often given of the persistence that would be induced by the normal operation of the transmission mechanism. The query makes sense in a book on macroeconomic theory, particularly so when we realize that most academic writings about European unemployment pretend to offer more than a diagnosis on a particular historical phase in a part of the world. They claim to provide a theory of unemployment persistence, a theory that would generally apply and would locate the source of the phenomenon in some widely prevalent features of the transmission, through the economic system, of the initial impact of shocks.

Consider for instance the article of O. Blanchard and L. Summers, "Hysteresis and the European unemployment problem" (op. cit.). The first section concerns "The record of persistent unemployment", where three European countries are compared with the US; the second section presents "A theory of unemployment persistence"; the third discusses "Empirical evidence on hysteresis theories", with again reference to European countries and the US; the fourth section answers the question "Is Eurosclerosis really the problem?". Similarly the survey article of C. Bean[129] is organized within a theoretical framework presented at the beginning without a discussion as to whether alternative theoretical frameworks would not have been enlightening for a full diagnosis about European unemployment: the author simply writes that competing approaches can be related within his framework, but attention of the reader is geared towards the elements that the framework selects as structuring the search for causes of persistence.

Most writings on unemployment persistence indeed start from the presupposition that attention should first bear on knowing how and how far do real wage rates adapt after macroeconomic shocks. It is even common to refer to a simple standard graph in order to support assertions about that adaptation. The terms in which the graph is explained differ somewhat, so we may wonder about its exact foundations. But we shall refrain from dwelling long on the differences

[129] C. Bean, European unemployment: a survey, *Journal of Economic Literature* (June 1994).

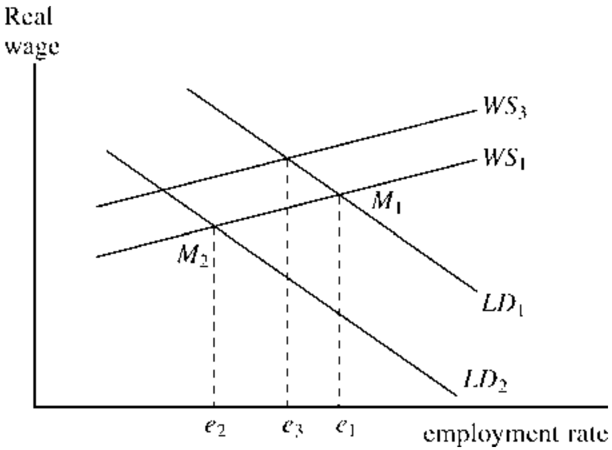

Fig. 8.

between authors. We shall rather take as an example an article by J. Elmeskov and M. MacFarlan[130].

Like in Figure 8, the employment rate $e = L/\overline{L}$ (or simply employment L in some presentations) is plotted as abscissa and the real wage rate w/p as ordinate. The initial situation is assumed to be at the intersection M_1 of two lines, respectively downward and upward sloping: the labour-demand schedule (also called by some authors the "price-setting schedule") and the wage-setting schedule. Implicitly or explicitly the first schedule is supposed to trace the decision of firms, whether they decide directly on their labour demand or on the price they offer on monopolistic markets (a price that will lead to a demand for their output, hence to their employment needs). The second schedule means that the determination of the real wage, by wage bargaining or otherwise, leads to a higher result when the level of unemployment is lower.

A permanent shock depressing the demand for labour (or at least a shock depressing it for a substantial number of quarters) means a downward shift of the labour-demand schedule from LD_1 to LD_2 and an increase in the unemployment rate from $1 - e_1$ to $1 - e_2$, the shift of the LD schedule being somewhat mitigated by a decrease in the real wage. (An increase in the unemployment rate could also follow from an upward shift of the wage-setting schedule from

[130] J. Elmeskov and M. MacFarlan, Unemployment persistence, *OECD Economic Studies*, No. 21 (Winter 1993).

WS_1 to WS_3.) As long as the labour-demand schedule remains in the low position LD_2, unemployment will persist if the wage-setting schedule does not shift downward. Elmeskov and MacFarlan distinguish between "unemployment hysteresis", when lasting unemployment has absolutely no effect on the position of the wage-setting schedule, and "unemployment persistence", when it has some effect, shifting *WS* downward, but slowly.

We may first wonder whether such an essentially static tool (Figure 8) may be appropriate for analysing persistence, which has a dynamic dimension: after a downward shift of the *LD* schedule, unemployment is meant to lead to two phases of decrease in the real wage rate, an immediate decrease along the *WS* schedule (the lag, if any, is left unspecified), and a subsequent slow decrease permitted by a shift of the *WS* schedule. We note also that, unless *WS* is drawn horizontal, hysteresis as defined by Elmeskov and MacFarlan means an immediate effect, but no subsequent effect; on reflection it looks like a rather strange property for the wage-setting process, whatever its microeconomic justification (wage bargaining, efficiency wage, ...).

More importantly, the graph conveys the idea that unemployment is due to the failure of wages to adjust downward sufficiently, or sufficiently fast, after an adverse shock on the demand for labour. We may be surprised that this idea is accepted off-hand without even a mention of all the earlier literature that disputed it, including *The General Theory*. Readers of this book know why the idea is disputable: whereas the demand for labour by a firm is in the medium run, other things being equal, a decreasing function of the real wage the firm has to pay, other things do not remain equal, particularly in the short run, when wages adjust downward. After such an adjustment, wage income and consumption decrease, demand for goods and services decreases, the downward revision in prices creates liquidity problem to indebted firms, which have to cut their demand for inventories, for labour and for investment goods. The macroeconomic dynamics is completely different from the one suggested by a focus on the two schedules of Figure 8.

Perhaps should we rather view the figure as a guide for medium-term equilibrium analysis, which may indeed seem relevant when and where unemployment is found to persist. But then a much more serious examination of the medium-run phenomena is required than is usually given in order to justify the standard graph of Figure 8. For instance, in the book of R. Layard, S. Nickell and R. Jackman (op. cit.), the graph is presented as serving for the determination of not the current unemployment rate but rather the "equilibrium unemployment rate". But the authors quickly identify the latter with the NAIRU and indeed turn most of their attention to the econometric estimation of price and wage equations of the type discussed here in Chapter 8. On his part C. Bean (op. cit.), in the eight pages of his Section 2 on “the theoretical background”, devotes most of the discussion to these price and wage equations and leads the reader to unconclusive results, which seriously question the whole approach.

More recently, writing about the “natural rate of unemployment”, O. Blanchard and L. Katz[131] insist on the relation of this rate to wage setting. They reproduce the standard graph as providing a representation of their views about the determination of the natural rate. They then speak of the “supply wage relation” instead of the “wage-setting schedule” and of the “demand wage relation” instead of the “labour-demand schedule”. More importantly they stretch the medium-run in the direction of the long-run and draw the *LD* as horizontal. They write: “If we focus for simplicity on the medium run – that is, the run over which firms can adjust all factors of production including capital – we can think of the real wage paid by firms as independent of the level of employment… The simplest interpretation of this relation is as a long-run labor demand curve, giving the real wage consistent with other input prices and the conditions that firms make zero pure profit”.

Accepting such a viewpoint here would be going much too far into the long run for our present subject, which is unemployment persistence. Let us rather go back to what was earlier explained in this book. The wage-setting schedule can be identified with “the wage curve” of Blanchflower and Oswald. But we remember the discussion in Section 5.6 of Chapter 7 concerning the conditions under which the macroeconomic demand for labour may be taken as negatively related to the real wage rate[132]. Indeed, a clear outcome of Chapter 7 is to suggest that unemployment diagnosis ought to distinguish between various possible combinations of market disequilibria and ought to identify which combination applies in each case.

Our discussion of short-run dynamics in Chapter 8, notably in Section 7.2, led us to related distinctions involving in particular, besides the unemployment rate, also the rate of utilization of productive capacities. Our examination of European unemployment in a moment, will find it appropriate to introduce, besides the real wage rate, also the real interest rate. Thus, there are good reasons to suspect that the dynamics of unemployment persistence depends on which combination of market disequilibria (quantity and price disequilibria) dominates the scene.

[131] O. Blanchard and L. Katz, What we know and do not know about the natural rate of unemployment, *Journal of Economic Perspectives* (Winter 1997).

[132] Note that in Figure 21 of Chapter 7 the real wage rate was plotted on the horizontal axis and employment on the vertical axis.

This dependence is systematically overlooked in the literature now discussed. This is why persistence has to be explained, mainly if not wholly, by rigidity of the real wage rate.

Overall, the foregoing discussion suggests that the literature of the 1990s gave too much weight to the idea that a framework for studying the causes of unemployment persistence could be provided by a graph such as the one exhibited by Figure 8: the phenomenon is really too complex to be captured through so simple a network. But of course the literature contains more than the disputable idea. Let us then look at its substantial content from two complementary points of view: first, how do various authors analyse the role of labour market rigidities; second, how do they explain European unemployment persistence.

2.9.5. *Rigidities as factors of unemployment persistence*

In order to explain real wage rigidity O. Blanchard and L. Summers (1986, op. cit.) propose a theory of wage bargaining in which only "insiders", but not other workers, would take part: after an adverse shock the newly unemployed workers would lose, more or less quickly, their insider status and the new smaller group of insiders would set the wage so as to maintain employment at its new lower level, with no consideration for the fate of the unemployed outsiders[133]. (When choosing its wage claim the group of insiders is assumed to take account of the supposed fact that the demand for labour is a decreasing function of the real wage, an assumption which is more likely to hold at the microeconomic than at the macroeconomic level.)

In order to try and demonstrate that their insider-outsider theory is explaining a higher unemployment persistence in three European countries (France, Germany, UK) than in the US, Blanchard and Summers look at two kinds of empirical evidence. They first examine what can be said about the size of the union sector, the role of unions in wage determination and the "membership rule" (who does the union represent). Overall this institutional evidence is inconclusive. Thus their discussion centers on estimations of wage and employment equations.

The most significant fit, according to the authors, relate real wage inflation to previous observation of employment. It may be written as:

$$[w_t - E_t p_t] - [w_{t-1} - p_{t-1}] = k + \alpha n_{t-2} - \beta n_{t-2} + \varepsilon_t, \qquad (149)$$

[133] In the discussion R. Hall remarks that, in order to so obtain sustained persistence, we need to assume asymmetry of the rule giving membership to the group of insiders: after a favourable shock newly recruited workers are slower to get insider status than the newly unemployed were to lose the status after an adverse shock.

where k, α and β are estimated parameters, ε_t is a serially correlated residual, w_t, p_t and n_t are the logarithms of respectively the wage rate, the price level and employment, finally $E_t p_t$ is the expected price level estimated by a simple linear adaptative rule with reversion to the mean. The fits of (149) and of variants of this equation are obtained from yearly data for 1953 to 1984, separately for each one of the four countries. The authors interpret the ratio β/α as a measure of persistence. It appears to be weak and poorly estimated in the US, but to stand in the range from 0.6 to 1.1 in the three European countries.

However, Blanchard and Summers do not take this result as sufficient to establish "whether the cause of hysteresis in Europe is unions or the sequence of adverse shocks which has caused high unemployment". Looking at other evidence, in particular at parallels between the European depression of the 1970s and early 1980s and the American Great Depression of the 1930s, they suggest that hysteresis in Europe might "be more the result of a long sequence of adverse shocks than the result of structural problems".

In their 1991 book R. Layard, S. Nickell and R. Jackman (op. cit.) focus on the unemployment-inflation trade-off and on estimations of price and wage equations in order to explain postwar unemployment in OECD countries (yearly data 1956–1985). They find large variations across countries in unemployment effects on wages, these effects being strongly inversely related to the duration of unemployment benefits and directly related to the proportion of small firms in the economy, as well as to the overall degree of inter-firm and inter-union coordination in wage bargaining. They also find that "unemployment hysteresis in wage-setting" is related to the proportion of long-term unemployed, with the duration of unemployment benefit being the key variable in explaining this proportion[134].

Similarly, a central piece in the 1993 article of J. Elmeskov and M. MacFarlan about unemployment persistence (op. cit.) is provided by estimates of wage equations for all countries concerned. The distinction between short-term and long-term unemployments is recognized as important for a proper charac-

[134] It is not the place here to enter into a detailed examination of the effects of unemployment-benefit schemes. These effects, which have been much studied by labour economists, seem to depend on detailed characteristics of the schemes, which changed through time and much differ between countries (see, for instance, J. Martin, Indicators of replacement rates for international comparisons, *OECD Economic Studies*, No. 26 (1996/1)). A very rough description of these effects consists in reporting that increases in benefits induce mainly increases in frictional unemployment and in participation in the labour force (this was indeed stressed in Subsection 3 of this section). Their role in wage formation is less obvious, but seems to well appears where the replacement rate is already high and when simultaneously wages tend to be "too high".

terization of wage inflation[135]. Increases in short-term unemployment appear to be definitely more powerful in containing wage rises than are increases in long-term unemployment. Typically an increase of one percentage point in the rate of short-term unemployment reduces the annual rate of wage increases by about one point also (with however large variations among country estimates); an increase of the same amount in the rate of long-term unemployment has an effect about three times smaller or even less. However, the fact that this last effect still appears significant in most countries is taken by the authors as evidence against strict hysteresis.

In a book published in 1993 A. Lindbeck devotes a chapter to "Mechanisms of unemployment persistence"[136]. After stating that "unemployment persistence is probably caused by several different mechanisms" he surveys a number of such so-called mechanisms which were proposed in the literature and concludes: "There are several plausible mechanisms of unemployment persistence. It is too early to say which of these mechanisms are the most important ones in the real world". In the survey we find factors which were already alluded to this section or earlier in this book, for instance the loss of human capital, recognized in Subsection 2.9.3 above, or the fact that, when classical unemployment has prevailed (see Chapter 7), capital shortage may soon appear and be maintained for some time after an increase in demand, which would have otherwise induced a larger and faster increase in employment.

Lindbeck gives a particular importance to the interplay between the turnover costs and the behaviour of insiders who, because of these costs, benefit from rents and have market power. The insider-outsider theory developed for this case by Lindbeck and Snower was briefly exposed here in Section 1.6 of Chapter 7. Using a simple model we then saw that, within margins, labour input was rigid. More generally Lindbeck argues that, along ordinary business cycles, employment fluctuates less if labour turnover costs are high than if they are low. But he further insists that the image opposing small to large fluctuations in employment is misleading if applied after a deep depression: in such a case labour hoarding will have been eroded, employment will start from about the same low level when the upturn will come, but it will increase more slowly if labour turnover costs are high than if they are low. Moreover, the wage-setting behaviour of insiders is likely to lead to higher increases in the wage rate if turnover costs are high; this will accentuate the persistence effect of turnover costs.

[135] Long-term unemployment is defined as the number of persons who remained unemployed for the preceding twelve months or more.

[136] A. Lindbeck, *Unemployment and Macroeconomics* (MIT Press, 1993).

Rigidities are more generally given a large responsibility in unemployment persistence by the well known *OECD Jobs Study*[137]. In order to promote employment and growth, countries are advised in particular: to increase flexibility in working-time patterns, to increase flexibility in wages and other labour costs, to reform job-security regulations or agreements, again towards more flexibility.

The exact justifications and implications of such an advice are debated among European economists, as can be seen in two articles published by the *Journal of Economic Perspectives* (Summer 1997). Horst Siebert writes under the title: "Labor market rigidities: at the root of unemployment in Europe", which is a brief but good summary. Stephen Nickell (Unemployment and labor market rigidities: Europe versus North America) concludes as follows: "It is clear that the broad-brush analysis that says that European unemployment is high because European labor markets are "rigid" is too vague and probably misleading. Many labor market institutions that conventionally come under the heading of rigidities have no observable impact on unemployment". Correctly presenting here the arguments of the debate would require a long piece that would take the reader into the study of labour market institutions and the econometric analysis of their effects.

2.9.6. *Causes of "European unemployment persistence"*

Before closing this unconclusive section it may be appropriate to still briefly focus on explanations of the European unemployment history between 1975 and the late 1990s. Persistence of European unemployment was much discussed in the literature, not only in the references already given earlier in this section but also in others[138], among which we shall select here the conclusions drawn by J. Drèze and C. Bean[139] from a set of parallel country studies conducted in the late 1980s.

Broadly interpreted, the various writings on the issue seem to agree on two conclusions. First, "there does not seem to be a single cause of the rise in European unemployment. Rather there have been a number of adverse developments". Second, "there is no sequence of adverse shocks that alone seems capable of rationalizing the persistence in European unemployment... Such persistence mechanisms seem to be far less pronounced in the United States"

[137] *The OECD Jobs Study: Facts, Analysis, Strategies* (OECD, Paris, 1994).

[138] See in particular A. Lindbeck, The West European employment problem, *Weltwirtschaftliches Archiv* **132** (1996), Helft 4, pp. 609–637.

[139] J. Drèze and C. Bean, Europe's unemployment problem: introduction and synthesis, in: J. Drèze, C. Bean, J.-P. Lambert, F. Mehta and H. Sneessens (eds.), *Europe's unemployment Problem* (MIT Press, 1991).

(both quotations from C. Bean, op. cit., p. 614). The second conclusion then leads to look for structural differences between the Western European and the American economies.

The author of this book does not fully agree with what appears to be close to a consensus view. He believes that *excessive influence is given by this view to structural factors*, because the part played by other factors, more congenial to macroeconomic thinking, has been underrated. He will then try and explain his own conclusions after a brief report of the synthesis drawn by Drèze and Bean (op. cit.).

(i) *The European Unemployment Programme* involved parallel econometric investigations made in nine European countries and the US within a common framework. Broadly speaking about the same model was used on annual data covering in most cases the years 1960 to 1985. The model, which will be more precisely defined in Subsection 4.1.3, determined jointly output and employment from demand and supply factors, which were themselves interdependent with prices, wages, unemployment and the degree of capacity utilization. The broad empirical regularities resulting from the Programme were summarized by Drèze and Bean in eleven points, among which we may select the following ones.

"Employment growth in Europe was curtailed by factor substitution at an average rate of some 2% per year until the late seventies, and 1% per year more recently. If that substitution had stopped when it became wasteful, the rise of unemployment in Europe could have been avoided". "The European economies are inflation-prone". "The main and nearly unique proximate determinant of output growth in the eighties has been effective demand". "Temporary demand stimulation through fiscal policy exerts only temporary effects, in within-sample-period simulations of the models. And the papers contain anecdotes about policy reversals induced by cumulated deficits".

The conclusions, which draw not only from the summary but also from the detailed results, state: "The empirical findings suggest explaining the contrast between the US employment miracle and the European persistent unemployment in terms of two proximate causes: (i) the wage formation process differs as between the two zones, in particular regarding the incorporation of measured productivity gains into real wages; (ii) whereas the proportion of firms where output and employment are demand determined grew markedly in all European countries in the late seventies and mostly remained high, that proportion is not trended in the US... The European wage-formation process makes non-declining employment dependent upon sustained output growth. In small open economies... when world demand or competitiveness are missing, fiscal policy alone is not a very effective instrument".

Two structural differences between Europe and the US are so stressed. The first one, namely specificities of the wage-formation process, is common to practically all explanations of European unemployment persistence. The second difference, the formation of aggregate demand and the potentialities of fiscal policies, is less unanimously recognized; when it is, the exact role given to structural factors varies from one author to another. Rather than commenting further on the literature about these differences, let us take some distance from it and look at the phenomenon from another point of view.

As Figure 7 shows, European unemployment began to rise in 1975. In a number of years since then the evolution was similar to that observed in the United States. Divergence concentrates on three periods: 1975–79 (the unemployment rate increased by 1.4 percentage point in Europe – the 15 countries of the present Union – decreased by 2.7 points in the US), 1982–85 (increase of 1.3 point in Europe, decrease of 2.5 points in the US), 1992–97 (increase of 1.4 point in Europe, decrease of 2.6 points in the US). Perhaps we could find a specific explanation for each one of these three periods, and perhaps the explanations would look different from those resulting from a search for structural disparities.

(ii) Most macroeconomists will agree that *the last period can be explained without reference to structures as usually understood*. The years 1992 to 1994 were marked in Western Europe by a specific strong debt-deflation that came at the end of a speculative boom in building and of a financial boom, partly engineered by the capital market deregulation of the late eighties, which had opened new credit opportunities to households and small firms. During the same years monetary policies remained quite strict in most European countries, with an annual real short-term interest rate equal on average to 5.2 per cent in the European Union as against 1.2 in the US[140]. Only in 1996 were the real short-term rates about in line at 3 per cent on both sides of the Atlantic, but then with business conditions which were so different! Long-term real rates were also higher in the European Union than in the US, the difference remaining at about 1 per cent through the years 1993 to 1996. Fiscal policies did not help either. Although the automatic stabilizers played their part between 1991 and 1993, the deficit of public administrations of the European Union then increasing from 4.2 to 6.1 per cent of GDP, a sharp reduction of this deficit was imposed in the following years (up to 2.4 per cent of GDP in 1997) notwithstanding the quite high rate of unemployment. The rationale of the European macroeconomic policy of those years could not be understood without refer-

[140] Nominal interest rate on average over the year *minus* rate of increase of the GDP deflator with respect to the preceding year.

ence to a political vision according to which the economic costs had to be born in order to give its best chance of success to the monetary unification, a long-range project indeed.

Macroeconomists and econometricians who had laboured in order to explain the 1970s and 1980s cannot be blamed, of course, because their explanation does not seem to be required for the 1990s. But the recent years show the relevance of traditional macroeconomic arguments. Moreover we may wonder whether full justice was given to a macroeconomic argument of a different kind, which had been put forward already around 1980 in order to explain the different evolutions of Western Europe and the US from 1975 to 1985, an argument which was seldom recalled more recently.

(iii) The decade 1975–85 was marked by stagflation and by the deep depression of the years 1980 to 1982. That particular depression was due to the conjunction of the second oil shock and of a sharp tightening of monetary policies meant to stop the two-digit inflation that was becoming common. In this respect conditions were similar in Western Europe and the US; among other things the real short-term interest rate jumped from a negative value in 1978 (−1.4 and −0.2 per cent, respectively in Europe and the US) to a fairly high value in 1981 (4.4 and 3.7 per cent). But the difference between the two regions was that recoveries were fast in the US after the two oil shocks, while unemployment kept increasing in Europe.

The difference was attributed by some economists to a particularly important "*real wage gap*" in Europe: real wage levels had drifted away, allegedly above the marginal product of labour at full employment[141]. Before commenting on the difficulties of the assessment and on its implication on unemployment, let us simply look at the labour share in GDP at factor costs, as it appears in the harmonized long series that are now available[142]. Whereas the labour shares were about equal in 1969 and 1970 on the two sides of the Atlantic[143] (in 1970, 74.0 per cent both in the US and in the Europe made of present European Union's countries), they drifted away progressively during eleven years, up to a maximum of five percentage points (in 1973: 72.4 per cent and 74.0 per cent; in 1979: 71.4 per cent and 75.4 per cent; in 1981: 71.9 per cent and 76.8 per cent). Six more years were required to bring again in line the labour shares at 72 per cent in 1987.

That such a macroeconomic difference could have had a responsibility in the relative employment performances was not so unlikely on a priori ground. The

[141] M. Bruno and J. Sachs, *Economics of Worldwide Stagflation* (Basil Balckwell, Oxford, 1985).

[142] *European Economy*, No. 65, Brussels (1998).

[143] However the difference between the European and the American labour shares had consistently exceeded 3 per cent during the years 1962 to 1967.

"wage gap hypothesis" stating this responsibility was studied and extensively discussed. But the subject went out of fashion soon after the publication the book of Bruno and Sachs (op. cit.). Looking at the literature about it, present observers would have to conclude, I am afraid, that the hypothesis was neither widely taken as proved nor widely taken as disproved. That such things could happen in our scientific discipline, about such a crucial issue, does not speak in its favour.

We must, however, recognize that the labour share can vary for other reasons than a changing difference between the real wage and the marginal productivity of labour. Thus defining a well-founded measure of the wage gap is delicate. Several attempts at doing so were proposed. They were surveyed in a thoughtful article by J. Helliwell[144], from which methodological conclusions can be drawn, besides other relevant hints and a support to the idea that, in some countries for a number of years, real wages were indeed excessive. The methodological issues are sufficiently important in the discussion of this subsection for justifying an incidental development here.

Any well-founded measure of the wage gap must rely on an explicit representation of the supply decisions taken by enterprises, in particular on a specification of their production function. Given the present stage of our macroeconomic knowledge, there are various possibilities for this representation and this specification; since they lead to significantly different results, no formula for measuring the wage gap is likely to get wide recognition. In other words, pretending to give a regular evaluation of the wage gap would be premature.

Moreover, published attempts at measuring the wage gap seem to all suffer from the same basic flaw. The wage gap has a natural definition; it is the difference between the current real wage rate and a value that would be consistent with a satisfactory level of the demand for labour by firms, assuming the demand for goods would otherwise be satisfactory. The hypothetical value of the real wage rate that is considered would permit an equilibrium of the labour market, while the current value is claimed to prevent it. Clearly, the contemplated equilibrium is not meant to be implemented in the short run, but after several years at least more often after a decade, so that productive capacities are themselves consistent with full employment. Now, all the measures so far proposed assumed the capital stock to remain fixed at its current level; they then considered a hypothetical increase of the demand for labour working on this capital stock The assumption so removes from the representation of the demand for labour its main proximate medium-term determinant besides the

[144] J. Helliwell, Comparative macroeconomics of stagflation, *Journal of Economic Literature* **26** (1988) 1–28.

demand for goods (see in particular Chapter 4, Section 3.5, and Chapter 7, Section 5.6).

Since the wage share does not provide an appropriate indicator and measures of the wage gap are still too ambiguous, we may think of looking for other evidence of a disequilibrium which, appearing in the system of prices and costs, would be an important cause of unemployment in the medium run. It so happens that we have such an evidence for the European Union and the period in question. It is provided by an indicator of the rate of return on productive capital, a rate of return which, in conjunction with the real rate of interest, directly bears on the profitability of production. With base 100 for the period 1961–1973, the indicator amounted to only 73 for the twelve years 1974–1985 (back to 90 for the decade 1986–1995, exceeding 100 from 1996 to 1998). With such a low profitability of production, particularly after 1980 when the real long-term interest rates went up to 4 and 5 per cent, it is not surprising to realize that European firms tended to donwsize their operations, hence their employment.

(iv) *The depression of profitability* was closely linked with the wage push. Hence, it is part of the phenomena studied by the European Unemployment Programme. The author of this book would like, however, to add two comments to the conclusions commonly drawn from this programme for the decade 1975 to 1985. In the first place, Drèze and Bean (op. cit.) analysed the "supply side" of the influence of wages on employment as operating through capital-labour substitution. It seems to me that the impact through the evolution of productive capacities was still more important; but econometric detection of the role of constraints on productive capacities is unfortunately difficult as we shall see in Subsection 4.1.4.

In the second place, Drèze and Bean explain the European wage push by an institutional difference with the US: productivity gains would be incorporated into wages much faster in Europe than they would be in the US. This explanation, coming from the fits of their wage equation[145], seems to miss the

[145] A result concerning this wage equation is not stressed by Drèze and Bean in their conclusions, but is relevant for a proper appreciation of unemployment persistence. The equation contains an error–correction term, which reflects a long-run constraint that short-run changes have to eventually satisfy. This term relates the real wage rate to the unemployment rate, or in some countries the labour share to the unemployment rate. But in all cases, in the US as well as in European countries, "the estimated sensitivity of real wages is too low to ensure that unemployment is substantially self-correcting in the face of adverse shocks" (p. 425 and 428 in Drèze and Bean). Since the sensitivity is weaker than that found by D. Blanchflower and A. Oswald (op. cit.), we may consider it as concerning the medium rather than the long run. But that is precisely the proper time dimension for judging the phenomenon of unemployment persistence.

importance of a factor which was not incorporated in the econometric fits on annual time series, but which I believe to have played a quite significant part during the seventies, even before the first oil shock. This might be called the "aspirations-expectations factor".

The European wage gap might indeed be seen as the result of inertia in the adaptation of expectations, norms, claims and concessions to a new phase in the socio-economic history of our countries. The two-decades-long phase ending in the late 1960s had experienced an exceptional economic growth, which had been both fast and regular. Households, workers, firms and governments had progressively taken the habit of reckoning on future increases in their real incomes. This habit had progressively shaped social norms on the wage rises that employers were expected to offer and trade unions expected to demand, on top of what social partners were considering as absolutely warranted, namely compensation for a creeping inflation, of which people were more and more aware. If this description is correct, the change from the 1960s to the 1970s has not to be explained by labour-market institutions which would be particular to Europe. It might rather be the change in the overall economic environment that was particular, since US growth had earlier been definitely slower and less regular.

I believe that the above vision could be documented by analysis of what was written at the time in a wide spectrum of papers, from those giving news in the public press to reports of large corporations and declarations of governments. If such an historical research project is undertaken, it might also give the opportunity of wondering whether, in some countries, the problem was not made worse by particular socio-political difficulties of the time: granting larger wage rises than the then prevailing norms may have been a way to cope with non-economic tensions.

In retrospect the hypothesis that the wage gap played a central part in the European unemployment problem of the 1980s may now appear corroborated: three of the four countries of the European Union which best recovered in the late 1980s and 1990s had an employment strategy giving a large role to wage moderation, a strategy that was early accepted and consistently followed in subsequent years. Those were Denmark, Ireland and the Netherlands, the fourth successful country being the United Kingdom.

The case of the Netherlands is particularly clear because of the well-known "Wassenaar agreement" reached in the autumn of 1982 by the central organizations of employees and employers, as the unemployment rate had peaked at 12 per cent. This was followed by years of substantial wage moderation, with government cuts in public sector wages, the legal minimum wage and related social security benefits. Between its peak in 1975 and 1989 the wage share in

GDP at factor costs decreased by 11 percentage points (as against 6 points on average in the European Union between the same years). The unemployment rate went down from 12 per cent in 1982 to 7 per cent in 1989 and 4 per cent in 1998 (corresponding figures for EU were 9, 8 and 10 per cent).

Emergence of similar social pacts in Denmark and Ireland, together with the fiscal consolidations to which we referred in Section 7.1 of Chapter 7, marked the beginning of definite improvements in employment. The decrease in the labour share was still more manifest in Ireland than in the Netherlands (25 percentage points between 1974 and 1998, as against 9 points on average for the European Union). After peaking at 17 per cent the Irish unemployment rate went down to 8 per cent in 1998. The evolution in Denmark was more moderate, changes in the labour share being close to those holding on average in EU; however, the profitability indicator evaluated by the Commission increased more quickly: by 25 per cent between 1974–85 and 1986–90, as against 20 per cent in EU; and again by 32 per cent to 1998 as against 24 per cent. From its maximum of 9 per cent in 1983 the Danish unemployment rate went down to hardly more than 5 in 1998. There is written evidence showing wide acceptance of a new social contract in both countries, implying fiscal consolidation, macroeconomic stability and a neo-corporatist wage policy.

The idea that a new social contract was accepted in the 1980s also applies to the UK, but it was of a different nature. Initially imposed by a new political majority it was for a time strongly opposed by trade unions, but progressively it received wider support. Across-the-board wage moderation did not play a particularly important part in the new strategy, which aimed more at restoring flexibility in labour markets and at promoting macroeconomic expansion by an independent monetary policy.

Overall, the literature seems to have overrated the role of structural factors as explaining the European unemployment problem because it underrated the part played by the wage gap and the depression of profitability. However the intention is not here to exempt structural factors from any responsibility. Along lines which are well argued in A. Lindbeck (1996, op. cit.), a full explanation has to reckon with many factors.

2.10. Debt deflation and the part played by financial factors

Since the beginning of this chapter we developed an overview of business cycles theory. This overview is still incomplete because it left aside the possible role of financial factors, which are often stressed in public debates and were the frequent focus of attention in older theories. No macroeconomic culture should ignore this aspect of business fluctuations. But we shall have relatively little to report here which would come from the recent literature. This is why we shall first devote a good deal of this section to the consideration of the book *Prosperity and Depression*, by Gottfried Haberler, first published more than sixty years ago[146]. The incursion into the history of economic thoughts will be justified by the subject in this particular instance. It will moreover give us an opportunity to reflect on what have been the changing and permanent elements in the views about business cycles since the beginning of the century. It will so play a similar role to the one of the reference made to R. Frisch (1933) when we introduced the stochastic vision of business cycles at the beginning of Part 2 of this chapter.

2.10.1. *Prosperity and Depression*

Haberler's book was one of the outcomes of a decision taken in September 1930 by the League of Nations to aim at coordinating analytical studies that were then undertaken on the problem of the periodic recurrence of economic depressions[147]. The first step was to survey existing theories in order to exhibit their common points, to bring out their divergences and to determine as well as possible the causes of observed divergences. This first step was commissioned to the young Professor Haberler. The book was a success and immediately became a reference in academic teaching. New editions were published with additions of important new material.

Here we shall refer to the first edition[148] because it better reflects the state of economic theories before the emergence of aggregate demand analysis, which was going to dominate for a few decades the study of business fluctuations. Indeed, the new material of later editions concerned mainly the flow of writings inspired by the appearance of Keynes' General Theory. This material is indirectly present elsewhere in our book where aggregate demand analysis has not been neglected.

The 1936 book was made of two parts dealing, the first, with the presentation of the various theories, the second, with the synthesis that Haberler drew from them. Our interest now concerns the first part. But the second is not negligible, particularly because of the clear statement of the methodological and substantial issues about which the author had to take position. Here we shall

[146] G. Haberler, *Prosperity and Depression* (League of Nations, Geneva, first published in French, 1936, and in English, 1937). More recent editions at Harvard University Press, 6th ed. in 1964.

[147] Another outcome was the econometric work of Jan Tinbergen, mentioned in the next part of this book.

[148] I worked actually from the text of the third edition in which the argument of the initial chapters was at places somewhat revised.

begin with three general remarks connecting the ideas conveyed by the book with those held today by specialists of business cycle theories.

Toward the end of the chapter introducing his second part, Haberler poses the question: is a general theory of business cycles possible? He then points to the fact that profound dissimilarities between business cycles exist, from one country to another, from one period to another. He further comments: "differences and dissimilarities between the various economic cycles are much greater than many theorists of business cycles seem to suppose". Indeed, such remains the situation nowadays. On the one hand, we saw in Section 2.2 that few "stylized facts" were common to all cycles. But, on the other hand, we tended to forget the observation when we dealt with theories: except in the previous section, we did not stress the idea that the nature of shocks may be more important in practice than the properties of the transmission mechanism; we did not insist on the importance of a diagnosis, in each case, about the dominant factors, which happen to shape the short-run evolution and which explain why a particular business cycle theory may then be more relevant than others.

According to Haberler dissimilarity between cycles is due to the plurality of causes. Theories differ mainly as to the importance they give to one particular factor or another, which is qualified as dominant or determinant. Some of the causes are economic, others extra-economic. A related distinction is between "exogenous theories", which give the main role to perturbations exterior to the economic system, and "endogenous theories", which emphasize reactions of the economic system, often even its fundamental instability. The method followed by Haberler for his review of theories proceeds from the presumption that it is more important to study how the economic system reacts to external impulses than to search for whatever regularity may exist in the occurrence of these impulses.

Let us note that, notwithstanding the multiplicity of causes and accordingly the multiplicity of theories, Haberler still believes in the usefulness of a "very general theory" which ought to deal with the most important features of business cycles and which would apply to a large number of cases without being void of any real substance. The aim of his synthesis was indeed to present such a theory, whose applicability would require specified conditions concerning the organization for money and banking, the system of prices and wages and elementary features of techniques of production.

Prosperity and Depression, and this is the second general remark here, gives a large role to market disequilibria. In the first part of the book the longest Chapter 3 concerns theories of overcapitalization. The short Chapter 4 is devoted to overindebtedness and to "horizontal disequilibria" (divergences in the evolutions of various outputs or primary inputs with respect to what will be required by the long-run evolution of either the commodity composition of consumption or the availability of primary resources), in contrast to "vertical disequilibria", which are divergences between the evolution of investment and what will be the long-run evolution of saving. Chapter 5 concerns theories of under-consumption. Analysis within chapters uses concepts such as "capital shortage", "manpower shortage", discrepancy between "the natural interest rate" and the market interest rate, "disappearance of profit margins", and so on.

In his second part Haberler has two main chapters, one on the cumulative processes concerning first, expansions, second, contractions, the other on the turning points, first, crises, second, recoveries. The chapter on cumulative processes studies such questions as why does output of durable goods grows faster in expansions that output of non-durables. The chapter on turning points may stimulate more the thoughts of disequilibrium-theorists. Haberler first distinguishes two categories of forces, accidental forces (read: random shocks) and "organic forces", which normally or necessarily result from the cumulative process that has been going on but counteract its perpetual extension. For each one of the two kinds of turning points the argument is made of three steps concerning (i) the immediate causes, (ii) the reasons why, along the previously prevailing process, the economic system had become less and less able to resist counteracting shocks, (iii) the generation of organic forces during the previous process. For instance, after an expansion, (i) a depressive shock on the demand for goods and services may or may not trigger a crisis, (ii) a progressive rigidification makes the economic system more an more sensitive to depressive shocks, (iii) organic forces are likely to have created serious desequilibria able to trigger or to accelerate a cumulative contraction[149].

The third general remark is to note here that the importance of monetary and financial factors appears more systematically in *Prosperity and Depression* than in our book. Haberler takes care of distinguishing two meanings of the word crisis, when used about economic matters. Crisis has a technical meaning in the theory of business cycles, namely the passage from a phase of prosperity to a phase of depression. But it also has a popular meaning: in the public use it refers to a financial crisis, which may involve fall of prices in stock-exchanges, acute financial difficulties, a succession of bankruptcies, mass withdrawals of deposits, even a bank panic. Haberler notes that some

[149] Some proponents of the theory of endogenous cycles may like to see the two main chapters of Haberler's second part as concerning, the first, the dynamics of a given regime (either expansion or contraction), the second, the abrupt shifts from one regime to the other. Clearly, the conception conveyed by Haberler's book is more qualified, but it is not radically opposed to the ideas of regimes and regime shifts.

downturns occurred without financial crises and some financial crises did not generate downturns. However, coincidence between the two was frequent.

In both parts of the book much attention is devoted to monetary and financial factors. Let us consider just the first part, i.e. the review of business cycle theories. We may put aside Chapter 7 concerning cycles in agricultural production and the short Chapter 6 devoted to "psychological theories", which mainly argue that, because of reactions of expectations to changes in the objective determinants of business cycles, fluctuations will be more violent than they would otherwise have been. About half of the remaining chapters of the review explicitly gives a crucial role to monetary and banking institutions.

Historians of economic thoughts would find much to report about the various aspects, identified by Haberler, concerning the nature of this role of financial institutions according to various economists. We shall not really enter here into such a report. This would exceed the limits assigned to our book, where the financial sector is not analysed. We had already many occasions to discuss the role of monetary policy. In a moment we shall examine the theory of debt deflation. Thus, it will be sufficient for our purpose in this section to recognize the importance ascribed to fluctuations in bank credits by theorists who wrote during the first decades of this century.

A main concern of economists at the time was to bring out the role of commercial banks in the determination of the quantity of money. When supplying credits to their clients, banks were providing liquidities. An isolated bank had not the direct power to much increase the money supply; but banks typically acted in the same direction one after the other, would it only be because their own liquidities depended on what other banks did. So the supply of credit had a part to play even in business cycle theories which were purely monetary.

The clearest and most determined proponent of the latter theories was R. Hawtrey who published a number of books between 1913 and 1937. He held that fluctuations in the quantity of money were reflecting the instability in the provision of credits, which was at times easy with low interest rates, at other times parsimonious with high rates, and that these fluctuations reflected on those in aggregate demand. In the later stages of expansions or contractions effects on prices and further on interest rates were felt; but these in turn led to procyclical variations in the velocity of money circulation. So a cumulative evolution persisted until banks revised their behaviour, perhaps because of a change in the policy of the central bank (this had to come under the gold standard).

For most specialists of business cycle theories the analysis had to be more complex because reactions in the allocation of resources also played a part. When credit was easy and interest rates were low, firms were said to shift too quickly towards more round-about techniques of production and to over-invest. The disequilibrium in the stock of productive equipments progressively built up and either triggered the depression after a time or worsened it when credit conditions were tightened. This "monetary theory of overcapitalization" particularly attracted Austrian economists such as L. von Mises, F. Hayek or F. Machlup; they could borrow from Wicksell the distinction between the natural rate of interest and the market rate.

2.10.2. *Financial crises and the perception of excessive indebtedness*

A thread running through a large part of the business cycle literature of the interwar period concerns the nature and role of financial crises. Careful readers of Keynes'own writings detect in them, including in the General Theory, the presence of this thread[150]. A number of modern macroeconomists are also now realizing the importance of the phenomenon; those are often the same macroeconomists who stress the role of financial determinants of investment, that of credit rationing more generally and of the credit channel in the transmission of monetary policy.

The author of this book has no better argument than an empirical one to give in favour of recognition of financial crises. A number of economists attempted to demonstrate by purely logical arguments that financial crises *had to occur* in capitalist economies. But the arguments so proposed are not very tight; they could hardly be found persuasive by inhabitants of other planets who would also receive from earth logical demonstrations that financial crises could not occur. However, the fact is that such crises *do occur*. Since they do, we should at least try to understand how they develop and what are their effects.

Observation shows that the initial development of a financial crisis in a country or a wider area has three interlinked features: (i) more and more banks, firms and/or households want to reduce their debts, hence to stop buying assets or to sell assets, (ii) more and more asset prices tend to fall, (iii) the net worth of more and more businesses and/or households is reduced. These three features were for instance identified by Irving Fisher as a sign of debt deflation[151]. At various places in his writings Keynes also stressed the importance of fluctuations in asset prices.

150 Among other references see those given in the two following articles concerning more generally our present topic: H. Minsky, The financial instability hypothesis: a restatement, in: H. Minsky (ed.), *Inflation, Recession and Economic Policy* (Wheatsheafs Books, New York, 1982), first published in 1978; M. King, Debt deflation: theory and evidence, *European Economic Review* (April 1994).

151 I. Fisher, *Booms and Depressions* (Adelphi, New York, 1932); I. Fisher, The debt-deflation theory of great depressions, *Econometrica* **1** (1933) 337–357.

Whereas these three features are inherent in the occurrence of financial crises, the shock, or even simply the phenomenon, that triggered it may vary. For simplicity we may say that ignition may appear on anyone of the interlinked features. A bubble of some asset prices may burst (see Chapter 8, Section 1.8). Business confidence may collapse, or households may downward revise expectations about their future income. Or still indebted agents may realize that they went too far and that their indebtedness has become excessive.

However, the fact that such occurrences at times generate depressions does not mean that they always do. For instance, the October 1987 crash in stock markets did not stop the boom, although it was expected to do so at the time. Business confidence or households expectations often vary with little more than temporary effects. As for the perception of excessive indebtedness it is usually ambiguous and hard to recognize before it generalizes.

The data necessary for the computation of debt ratios exist, relating for instance, gross debt of firms or households to their value added, income, or net asset. But periods of excessive indebtedness are seldom easy to identify by inspection of time series of such ratios. For instance, the debt ratio of households increased during the 1980s in most Western European countries. But the phenomenon was a natural consequence of financial deregulation; it was not a matter of concern for business analysts before direct evidence was available showing that more and more households wanted to deflate their debts.

Over the business cycle the actual evolution of debts is moreover subject to the same phenomenon as the actual evolution of inventories: it is in part involuntary. When the boom slows down, expectations of a number of firms and households with respect to their incomes are disappointed whereas previous commitments have to be met. For instance, while commenting on the British debt deflation of the 1989 to 1992 period, M. King (op. cit.) also gives figures showing that private debts increased during those three years (by 7 percent in real terms).

2.10.3. *The depressing effect of debt deflation*

But, if widespread, the intention to reduce debts is a strongly depressing factor of aggregate demand. The phenomenon is most simply analysed in real terms, although induced effects through changes in incomes, prices and interest rates interfere in the exact determination of the results, in the same way as they do after autonomous changes in aggregate demand. The main point to note is the fact that the marginal propensity to spend from wealth is greater for debtors than for creditors. The reason is well explained in the following quotation of J. Tobin[152] (p. 10).

> "The population is not distributed between debtors and creditors randomly. Debtors have borrowed for good reasons, most of which indicate a high marginal propensity to spend from wealth or from current income or from any liquid resources they can command. Typically their indebtedness is rationed by lenders, not just because of market imperfection but because the borrower has greater optimism about his own prospects and the value of his collateral, or greater willingness to assume risk and to die insolvent, than the lender regards as objectively and prudently justified. Business borrowers typically have a strong propensity to hold physical capital, producers' durable goods. Their desired portfolios contain more capital than their net worth – they like to take risks with other people's money. Household debtors are frequently young families acquiring homes and furnishings before they earn incomes to pay for them outright; given the difficulty of borrowing against future wages, they are liquidity–constrained and have a high marginal propensity to consume"[153].

There does not seem to be so much econometric evidence on this assumption of a higher marginal propensity to spend of debtors, an assumption which however seems quite likely to hold. For households, where it would be easier to test than for firms, data on total indebtedness and net worth are lacking. But M. King (op. cit.) points to an important difference between the reactions of British households who owned their dwellings, depending on whether they had or not a mortgage debt: the decline in expenditures of these households between 1989 and 1991, at the time of the British debt-deflation, concentrated on those who had a mortgage.

The role of the assumption in explaining the aggregate depressing impact appears in the following argument. The perception by debtors that their debt is excessive must come from a reassessment of their expectations about the real burden of the debt, most probably because of a reassessment of their price ex-

[152] J. Tobin, *Asset Accumulation and Economic Activity* (Basil Blackwell, Oxford, 1980).

[153] We note in passing that this quotation is drawn from a chapter entitled "Real balance effects reconsidered", in which Tobin discusses whether the decrease in the price level that may be induced by a demand slack mitigates this slack. To the "Pigou effect" asserting a positive impact via the increased purchasing power of nominally denominated assets, Tobin adds the "Fisher effect" leading to an overall negative impact via the increased real burden of debts and the redistribution of net real assets from debtors to creditors. We discussed the phenomenon in Section 3.3 of Chapter 7, using a slightly different presentation.

pectations. But the same revision of price expectations should lead creditors to revise upward their income expectations. If the marginal propensity to spend of creditors was as high as that of debtors, the positive impact on creditors' consumption would exactly compensate for the negative impact on debtors' consumption. But the difference between the two marginal propensities explains the negative impact on aggregate households' consumption.

Another significant factor in debt deflations comes from the induced decrease in the probability that loans will be repaid. This decrease leads to an increase in risk premia charged on new loans. It also leads lenders to choose tighter norms in the amount of new credits granted to borrowers. Thus, credit rationing increases and some debtors become unable to repay loans, which would have created no problem under normal credit conditions. In this new financial environment, both borrowers and lenders are reducing the size of their real operations.

About the depressing effect M. King (op. cit.) gives further evidence coming from cross-sections of countries and of UK regions. In particular, considering a sample of ten OECD countries, he regresses the short-fall in consumption growth (defined as the difference between the growth of consumption in 1989–92 and its average growth in 1974–89) against the change in the ratio of household debt to GDP between 1984 and 1988. He then finds a significantly positive regression slope: the short-fall in consumption was systematically higher where household debt had previously increased faster.

2.10.4. *A credit cycle?*

As we saw, a depression triggered by the perception of excessive indebtedness is likely to cause in the short-run involuntary increases in debts. This means that excessive indebtedness will persist for some time. Thus, the dynamics of indebtedness is an interesting subject to study, notably in the same spirit as we studied unemployment persistence. Knowing better this dynamics might conceivably lead to giving justification to the idea of cyclical fluctuations in indebtedness, hence to the expectation of recurrent occurrence of debt deflations, hence further to explanations of business cycles which would give the main role to what was sometimes called "the credit cycle".

Knowing that debt-deflations are more generally depressive does not suffice to give a full theory of the business cycle. Irving Fisher presented his analysis as providing a "theory of great depressions". How far did he achieve this aim? Commentators at the time were not favourably impressed. According to G. Haberler in his survey published a little later (op. cit.), Fisher discussed an interesting aspect of the business cycle generated by overcapitalization, but he did not provide a complete new theory of the cycle, because he did not explain how the accumulation of debts could cause the collapse of the boom. Haberler agreed that indebtedness was worsening the depression because lower prices increased the real burden of debts, hence the debt deflation. But he held that excessive indebtedness was the consequence of overcapitalization, which was just as unsound whether financed by retained earnings or by borrowing. Some readers may think that Haberler, who did not mention changes in expectations in his analysis of Fisher's contribution, was too much thinking in terms of deterministic theories of the business cycle and had not really drawn the conclusion of Frisch's plea for a stochastic conception of business cycle theories. Other readers may note that Haberler did not anticipate the asymmetric-information arguments of some modern theorists of financial economics.

Be that as it may, modern theorists may also think that Fisher did not provide a complete model of his theory. Actually, Fisher identified other factors than the three listed in the Subsection 2 above, factors which according to him also characterized the cyclical tendencies associated with the expansion and contraction of debts: (iv) real interest rates rise in debt deflation, in relation with both the fall in asset prices and an increase in the perceived volatility of these prices, hence an increase in risk premia; (v)–(vi)–(vii) output, profitability and business confidence are then depressed; finally (viii)–(ix) the amount of bank credit, the quantity of money and the velocity of circulation decrease. The downward spiral stops when the action of debtors in cutting back consumption and investment in order to reduce debts are offset by the action of creditors who are able to increase spending.

Considering the variety of models now available in our tool-box, we can imagine and model societies in which credit and debt behaviours will induce business cycles, particularly endogenous cycles which can in some cases be associated with the multiplicity of stationary equilibria (see Section 2.6 above). This is done by M. King (op. cit.) who considers a society in which some risks are uninsurable. There are two types of agents with different random future endowments: one type has to be debtor, the other creditor. Attention is focussed on aggregate demand as a function of the expected future value Eq of an illiquid asset. The function is increasing for low values of Eq. Multiple equilibria are then possible, as well as endogenous cycles.

C. Azariadis and B. Smith[154] come closer to the microeconomics put forward by proponents of the credit view. They consider an overlapping generation economy with intragenerational heterogeneity and adverse selection in the capital market, where loan repayment is uncertain. In order to cope with

[154] C. Azariadis and B. Smith, Financial intermediation with regime switching in business cycles, *American Economic Review* (June 1998).

the adverse selection problem, loan contracts specify both the interest rate and the amount lent. Production requires capital, hence borrowing. In their economy (which is particular in a number of respects, as well as King's economy was) there is a range of values for the current capital stock, or equivalently a range of factor prices, which make the full-information allocation of credit incentive compatible; the law of motion of the capital stock then coincides with that found under perfect information and competition. However, there is also a range of values of the capital stock for which incentive constraints must be binding and credit be rationed; then capital formation and output are lower than would be observed under complete information. Finally, there is a range of values of the capital stock for which the two regimes are feasible: a Walrasian regime equivalent to that resulting from perfect information, or a regime of credit rationing with a low real rate of return on savings; which regime holds depends on savers' beliefs about the regime in which they are; a multiplicity of regime transitions can be observed. There are, however, conditions under which there is no equilibrium path consistent with unchanging savers' expectations: all equilibria display regime transitions. Again, under some conditions there exist endogenous cycles, i.e. deterministic perfect-foresight equilibria in which there are m periods of expansion (in a Walrasian regime), followed by n periods of contraction (in a regime of credit rationing).

Pretending to use the models proposed by King and Azariadis–Smith for comparative assessment seems to be out of question because any attempt at calibrating their parameters would look wholly artificial. The kind of cycles to which they may lead certainly overemphasizes the real importance of regime switches in which sharp revisions would occur in expectations, and possibly in the importance of credit rationing. They rather aim at suggesting how persistence in debt deflation or in the expansion of credits could be explained. They may be viewed by some as contributing more to a theory of financial instability than to theories of business cycles.

The next step should probably be empirical. It would have to look for indicators of financial expectations, of asset prices (see Section 1.8 of Chapter 8) and of the extent of credit rationing. It would have to systematically study how such indicators change along business fluctuations. The econometric literature probably contains already a number of contributions which would be relevant within such a research project[155].

[155] See in particular S. Wadhwani, Inflation, bankruptcy, default premia and the stock market, *Economic Journal* (March 1986).

2.10.5. *The conclusion of a wise man*

Ending this long Part 2 which ran over so many pages in order to survey business cycle facts and theories, we can find no better conclusion than to refer the reader to an article of V. Zarnowitz[156]. The author, who over half a century devoted his all professional life to the empirical study of business cycles, was asked to contribute to a symposium in the *Journal of Economic Perspectives* about the US "new economy" intellectual fashion claiming that the boom-bust business cycle was over. After dismissing in three pages such a euphoria reminiscent of the 1920s or 1960s, Zarnowitz devoted his article to present his views about "the fundamental yet unresolved issue of the underlying causes of business cycles".

His message is well summarized in the following quote: "Business cycles are far from being all alike; their symptoms and causes differ over time and across economies. But... at the center of business cycles are interacting movements in business profits, investment and credit. Their rise are cumulative and mutually reinforcing, and so are their declines. Moreover, the three factors play critical roles in explaining what happens at the downturns and upturns of total business activity in market economies. Fluctuations of profits, investment and credit have been a common feature of business cycles under very different monetary and exchange regimes, through inflations and deflations, in the 19th and 20th century, in Europe and Asia as well as in America" (p. 73).

Here in a nutshell we find the list the three main aspects of the business cycle phenomenon and the three main pillars of its explanation: credit, stressed in this section; investment, stressed in the first part of the chapter; and profits, the main manifestation of price disequilibria, which would certainly have also been stressed in this book if there had been to report more systematic work done about them in macroeconomics during the past decades.

[156] V. Zarnowitz, Theory and history behind business cycles: are the 1990s the onset of a golden age?, *Journal of Economic Perspectives* (Spring 1999).

3 Structural Macroeconometric Models

A survey of the various phenomena shaping business fluctuations, like the one made up to now in this chapter, leads to a large number of conclusions, often not so firmly established, sometimes mutually contradictory as to the direction of effects to be expected from some exogenous shocks. Thus, the dynamics of short-run economic movements appear as produced by a complex set of causes which act simultaneously. In order to model these movements we cannot be content with an analytical and qualitative approach, even though such an approach may often suffice for explaining the part played by one cause taken in isolation. We also have to adopt an overall and quantitative approach, suitable for a simultaneous representation of various phenomena and of their respective intensities. In academic research the approach can have a variety of applications, each one meant to offer a particular perspective on the dynamics in question. Indeed, when for instance we surveyed the research of the RBC movement, we saw quite a few such applications of the same methodology, meant in each case to provide an overall and quantitative representation, various representations complementing each other.

We are now turning our attention mainly towards applications to actual countries, for economic forecasting or policy analysis. In many cases such applications also require overall and quantitative models, but they are less tolerant of the diversity of perspectives: users do not like eclectic answers and cannot make anything of ambiguous answers. Thus, we shall focus our discussion on cases in which a macroeconometric model is built for a country in order to directly serve for a wide range of applications by the teams in charge of forecasting short-run trends or of advising governments about their macroeconomic policy.

The idea of building such models arose in the 1930s[157]. It was first implemented by the Dutch economist Jan Tinbergen who devoted much of his efforts to it between 1935 and 1950. Continuing in the same spirit, a whole school of quantitative economists specialized in constructing national econometric models which meant to explain short-run changes in economic activity. The most representative and most active scientist of this school was Lawrence Klein[158], who promoted implementation of models in many countries.

When the econometric methodology for the quantification of such models (and of others also involving simultaneous equations) was studied, already in the 1940s, the phrase "structural models" was introduced. It still prevails today, particularly when the methodology, which was subsequently extended so as to cover also applications of the models, is being compared with alternative approaches to quantitative macroeconomic assessments. Notwithstanding a good deal of research on such alternative approaches, structural models remain the main instrument actually used at the aggregate level by teams giving public information on business trends or advising governments.

Presentation of structural models in a book such as this one is now confronted with two difficulties, a major and a minor one. The main difficulty comes from the fact that some academic economists, even among the most renowned, have chosen to discredit structural models when they were addressing their readers and students. The line of argument has been hammered so much that it is reproduced faithfully and uncritically by authors who should be better informed about the actual scientific and practical conditions in which the work of macroeconomic forecasting and analysis operates.

The author of this book cannot overlook this ideological background; but he must expose, as well as he can, the methodology of structural macroeconometric models, the valid criticisms raised against it, the possible adaptations of the methodology and the potentialities of alternative methodologies. Faced with the difficulty of his task in the present intellectual environment, he has deliberately chosen to write this part of his book essentially as he would have done

[157] The history of business-cycle macroeconometric models received much attention in M. Morgan, *The History of Econometric Ideas* (Cambridge University Press, 1990).

[158] The "Klein–Goldberger model" can be considered as typical of the outcome of the first generation of macroeconometric models. It has been studied a great deal and served to many teams as an example when the first models for macroeconomic forecasting and policy analysis were built. See L. Klein and A. Goldberger, *An Econometric Model of the United States*, 1929–1952 (North-Holland, Amsterdam, 1955).

twenty-five years ago, although doing justice to more recent contributions that fit in the then prevailing framework. The next part of the chapter will be devoted to all the developments which, during the last two decades, questioned the worth of the framework and attempted to replace it or which, in some cases, just wanted to amend or to extend it.

The minor difficulty is to know how to show, in a few pages, what are the common characteristics of models which differ from one another and which contain more variables and equations than can be reported, in some cases many more. Our book must simplify and cannot give a complete account of what structural macroeconometric models are and have been[159]. The difficulty is, however, mitigated by the fact that a natural entry for our report exists, suffices for a good introduction to the methodology of structural models, and permits an appropriate baseline for quick references at the many parts of these models that cannot be displayed. This baseline transposes aggregate demand analysis from the theoretical models, examined in this book at various places starting with Chapter 7, to quantitative models anchored on the national income-expenditure accounts.

The first section here will indeed present such a baseline mini-model. It will serve as a reference for Section 3.2 formalizing the system of equations exhibited by any linear structural model. Section 3.3 will sketch the methodology for the quantification of the parameters contained in the initial specification of a model. Sections 3.4 and 3.5 will concern respectively the practice and the analysis of model-based macroeconomic forecasts. Sections 3.6 and 3.7 will be geared to policy analysis and will introduce a number of ideas and tools, some of which really bear on a wider domain than the use of structural models. A short last section will take stock of what was learned during fifty years of experimentation, study and application of structural models[160].

3.1. A representative mini-model

Many economists[161] noted that, from the 1950s to the early 1970s, most structural models shared the feature to constitute extensions of roughly the same

[159] Readers may find much more in the available literature, for instance, in R. Bodkin, L. Klein and K. Marwah (ed.), *A History of Macroeconometric Model-Building* (Edwar Elgar, Cheltenham, 1991).

[160] Books exist which deal much more fully with the main aspects which we shall survey here. See, for instance, R. Fair, *Specification, Estimation and Analysis of Macroeconometric Models* (Harvard University Press, 1984).

[161] See, for instance, M. Deleau, P. Malgrange and P.-A. Muet, A study of short-run and long-run properties of macroeconometric dynamic models by means of an aggregative core model, in:

simple Keynesian model, which could be used as an introduction in teaching. Depending on the exact list of variables and on the form of some structural equations, there are many variants of this core model. Here we shall refer to a specification due to M. Deleau, P. Malgrange and P.-A. Muet (op. cit.) and call it "the DMM model".

Things are even made easier for us by the fact that, within the core model, we can usually identify a real sub-model which functions almost independently in the short run. Studying this sub-model, ignoring the feedbacks from the rest of the model, then gives a good approximation of the first effects of most contemplated changes in exogenous variables. It is therefore worth introducing the core model in two steps, as we shall do here for the DMM-model. First, we shall specify a system with a hierarchical structure, in which a block of "real" equations can be dealt with separately, since the variables determined in the other blocks do not feature in it. In a second step we shall study the various channels through which these "non-real" variables are affecting real variables, that is, the various types of feedbacks that the DMM model contains: to its real block from its other blocks.

3.1.1. *The real block of a hierarchical model*

In a Keynesian spirit our core model explains the formation of aggregate demand for goods and services, linking this demand to the level of output as well as to those variables which characterize the degree of utilization of factors of production. If we ignore the influence of the price system and of nominal variables, we can easily isolate a real sub-model. We are then usually led to something fairly similar to the multiplier-accelerator model, which we examined at the beginning of the chapter. Household demand (consumption) and the demand of firms (investment) are included, as is also the demand of foreign agents (exports); production, employment and imports adapt to these according to well specified equations.

For example, an approximation of the real sub-model of the DMM model can be easily defined. It has the natural structure illustrated by Figure 9. Three exogenous variables feature explicitly: exports EX, government demand G and labour supply $\overline{L}$ (in order to simplify the notation the index t for the period is not shown). Nine endogenous variables, for which notations are too familiar to need reminders, are determined by nine equations, of which the following five are accounting identities.

P. Lagrange and P.-A. Muet (eds.), *Contemporary Macroeconomic Modelling* (Basil Blackwell, Oxford, 1984).

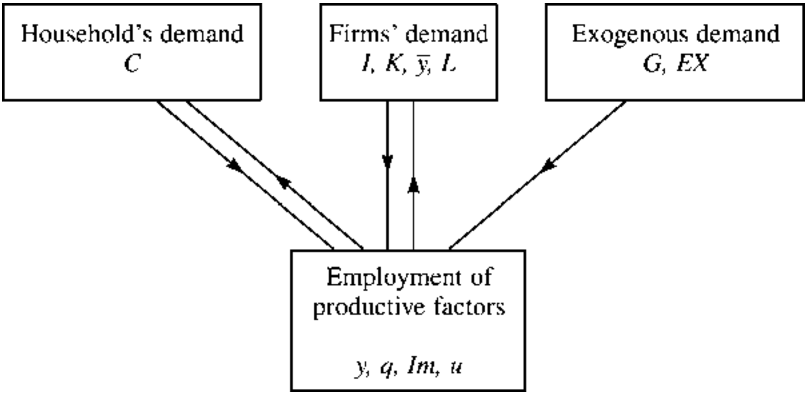

Fig. 9.

– Equilibrium holds on the market for goods and services:

$$y + Im = C + I + G + EX. \tag{150}$$

– The capital stock K_{+1} in the following period is determined by investment, once depreciation at rate δ is taken into account:

$$K_{+1} = I + (1 - \delta)K. \tag{151}$$

– Productive capacity is defined as:

$$\bar{y} = K/b, \tag{152}$$

where b is a constant[162].

– The rate of capacity utilization and the unemployment rate are defined as:

$$q = y/\bar{y}, \qquad u = 1 - \frac{L}{\overline{\overline{L}}}. \tag{153}$$

– The four behavioural equations are simple and familiar.

[162] Actually the DMM model allows b to change slowly and so takes technical progress into account. For the sake of simplicity we rule out here all terms corresponding to such trends.

The consumption function

$$C = (1 - \lambda_5)cy + \lambda_5 C_{-1} \tag{154}$$

includes two constants c and λ_5, the second one characterizing the lags in adaptation[163]. In order to define the demand for investment and labour we start from the expected level of production y^*, which results from adaptive expectations[164]:

$$K_{+1} = (1 - \lambda_3)by^* + \lambda_3 K. \tag{155}$$

We note in passing that, taking (151), (155) and (156) into account, and using here the lag operator L (not to be confused with labour input), we can write:

$$I = \frac{(1 - \lambda_2)(1 - \lambda_3)[1 - (1 - \delta)L]by}{(1 - \lambda_2 L)(1 - \lambda_3 L)}, \tag{156}$$

which is one of the possible specifications of the accelerator.

Likewise, the demand for labour is given by:

$$L = (1 - \lambda_4)ny^* + \lambda_4 L_{-1} \tag{157}$$

and the demand for imports by:

$$Im = ay^{e_1}q^{e_3}, \tag{158}$$

where a, e_1 and e_3 are positive constants.

At this stage we could examine the cyclical dynamics that this real sub-model would have if, as is done for structural econometric models, its coefficients had been estimated from the time series of the country concerned. We shall examine this question later, for an approximation of the real sub-model, for the complete DMM model and for macroeconometric models which are actually used. We can already suspect that the results will be fairly similar to those found in Section 1.2 for the multiplier-accelerator model.

[163] For most parameters we use here the notation in the article by M. Deleau, P. Malgrange and P.-A. Muet already quoted, so that it will be easier for the reader to use this article, which contains useful details not mentioned here.

[164] Here and in other equations we have replaced multiplicative expressions by additive ones, for the sake of simplicity.

Note further that, compared with the macroeconometric models in use, the most important simplification made here is that the sub-model ignores inventory accumulation or decumulation, whether voluntary or involuntary. Other simplifications come from the high level of aggregation, all goods and services being regrouped into just one.

3.1.2. *Other blocks of a hierarchical model*

Any model claiming to cover the whole economic field, even if in a very sketchy manner, must provide a representation of the price system and of its changes over time. In the DMM model we can, moreover, break down the block concerning this system into two parts: a "price-wage" block and a "cost of capital" block. To these we have to add a "monetary block".

The price-wage block uses tension indicators, which are the unemployment rate u and the rate of capacity utilization q, as well as the productivity of labour y/L, as an element of the cost of labour per unit of output wL/y. The block simultaneously determines the price level p, wage rate w and the expected rate of inflation g^e (under adaptive expectations):

$$g^e = (1 - \lambda_6)\frac{p - p_{-1}}{p_{-1}} + \lambda_6 g^e_{-1}. \tag{159}$$

The equation for the wage rate retains an immediate adaptation to the price level and a Phillip effect:

$$\frac{w}{w_{-1}} = g_1 + g_2\frac{p}{p_{-1}} + g_3 u, \tag{160}$$

where g_1, g_2 and g_3 are three constants. The equation for the price level includes the unit cost, the expected rate of inflation and the rate of capacity utilization. After linearization it can be written as:

$$p = (1 - \lambda_7)\left[f_1\frac{wL}{y} + f_2 q\right] + \lambda_7\big(1 + g^e\big)p_{-1}. \tag{161}$$

Once the real block is solved, the three equations (159) to (161) determine p, w and g^e. They do not have the exact form which we studied in Chapter 8, Part 6; but the factors involved are approximately the same.

Finally, the monetary block contains a very simple equation concerning the user cost of capital r/p:

$$r/p = i - g^e + \delta. \tag{162}$$

This is hardly more than a definition. But it includes the interest rate i, which is also determined in the monetary block. The demand for money assumes adaptation lags, but apart from this it is familiar:

$$M = (1 - \lambda_9)kpyi^{\mu_2} + \lambda_9\bigl(1 + g^e\bigr)M_{-1}. \tag{163}$$

Money supply is a function of an exogenous "monetary base" $\overline{M}$ but increases with the level of the interest rate:

$$M = \overline{M}i^{\mu_1} \tag{164}$$

(k, μ_1 and μ_2 are constants). Once the real block and the price-wage block are solved, Equations (162) to (164) determine the three last endogenous variables, which are r, i and M.

Thus the logical structure of the hierarchical model is very simple. It is so simple that we cannot be content with just this, either for explaining cyclical fluctuations or for thinking about macroeconometric models. Even at this preliminary stage we have to take account of feedbacks from prices and the quantity of money to the real temporary equilibrium. To sum up, there has to be more "integration" between the different blocks.

3.1.3. *An integrated model*

The DMM model retains most of the channels of integration which were mentioned in Sections 1.5 to 1.7 of this chapter. In addition it takes account of the elasticity of international trade with respect to the domestic price level, the exchange rate and foreign prices being implicitly taken as exogenous.

The equation for imports (158) is thus replaced by:

$$Im = ay^{e_1}p^{e_2}q^{e_3}. \tag{158$'$}$$

Whereas exports are no longer considered as exogenous but are determined by:

$$Ex = p^{e_4}\overline{Ex}. \tag{165}$$

An effect of changes in income distribution appears in the consumption function, Equation (154) being replaced by:

$$C = (1 - \lambda_5)\frac{c}{1 - \alpha\theta}(y - \theta P) + \lambda_5 C_{-1}, \tag{154$'$}$$

where α and θ are two coefficients, whereas P is the share of real income which does not go to labour:

$$P = y - \frac{w}{p}L. \tag{166}$$

Finally, the last element of integration, changes in relative prices affect firms' decisions. The DMM model uses a Cobb–Douglas production function and adaptive expectations to define desired capital K^*_{+1} and desired employment L^*, Equations (155) and (157) being then replaced by:

$$K_{+1} = (1-\lambda_3)K^*_{+1} + \lambda_3 K, \tag{155$'$}$$

$$L = (1-\lambda_4)L^* + \lambda_4 L_{-1}. \tag{157$'$}$$

The variable ω representing the expected relative cost of capital with respect to labour is determined by:

$$\omega = (1-\lambda_1)\frac{r}{w} + \lambda_1\omega_{-1}. \tag{167}$$

It thus depends on equilibrium of the money market, as well as on the time path of wages and prices. Desired capital and desired employment are specified as:

$$K^*_{+1} = by^*\omega^{-\beta}, \tag{168}$$
$$L^* = ny^*\omega^{\alpha}, \tag{169}$$

where α and $\beta = 1-\alpha$ are meant to be the exponents of L and K in the Cobb–Douglas function (α already figures in the consumption function).

Productive capacity is itself dependent on the real expected cost of labour σ:

$$\sigma = (1-\lambda_8)\frac{w}{p} + \lambda_8\sigma_{-1} \tag{170}$$

and Equation (152) is replaced by:

$$\bar{y} = \frac{1}{b}K\sigma^{-\beta}. \tag{152$'$}$$

With regard to the various factors considered in business cycle theories, the DMM model ignores in particular firms' financial resources, credit rationing and wealth effects. Most of the structural macroeconometric models have included these factors, whereas changes in relative costs were often made to play a smaller role. Obviously, in addition to these differences, many others exist, concerning the level of aggregation, the lag structure and more generally the specifications of the whole set of equations. Nevertheless, conceptions which shaped the construction of the DMM model are representative of those which prevailed in the elaboration of the much more complex models which were built for actual economies.

3.2. A general form for linear models

Although simple in comparison to actual structural models, DMM is already complex enough to suggest the challenge we should face if we would want to go deep into the discussion of methodological issues. Indeed DMM combines a number of features which are difficult to simultaneously embrace in any attempt at a general treatment of the properties of alternative methods for estimating or using structural models. We have two categories of variables, exogenous or endogenous, with more than a few variables in each category. We have two kinds of equations, accounting identities and behavioural equations. Many parameters appear in behavioural equations. In any equation only a subset of all variables is present. But values taken by the same variables in different periods may occur (see Equations (151) and (154′), for instance). Some equations are non-linear, like (158′) or (166). Finally the whole set of equations exhibits none of these symmetries which make other models more easy to handle (remember input–output models).

The natural response of methodologists to such a challenge is to decompose the difficulties and to long refrain from considering them all simultaneously. For instance, a number of issues are clarified by a first examination in which the presence of lagged values is ruled out from the system of equations (a simplification which, however, alters dynamic properties). Still more frequently all equations of the system are assumed to be linear. We shall, at this point and often in future developments, accept this second simplification, which may be justified as a first approximation for differentiable systems. Later in Section 4.1 we shall refer to some non-differentiable systems, which are relevant with respect to some issues concerning the macroeconomics of business fluctuations.

Let us then consider the following formalization. A number m of exogenous variables x_{it} $(i = 1, 2, \ldots, m)$ concern period t (most often a year or a quarter). There are also n endogenous variables y_{jt} $(j = 1, 2, \ldots, n)$ and n linear equations:

$$\sum_{j=1}^{n} \sum_{\tau=0}^{h} b_{kj\tau} y_{j,t-\tau} = \sum_{i=1}^{m} \sum_{\tau=0}^{h} a_{ki\tau} x_{i,t-\tau}, \quad k = 1, 2, \ldots, n \tag{171}$$

in which the $n(n + m)(h + 1)$ coefficients $b_{kj\tau}$ and $a_{ki\tau}$ are fixed. We can equivalently write the system in the following matrix form:

$$\sum_{\tau=0}^{h} B_\tau y_{t-\tau} = \sum_{\tau=0}^{h} A_\tau x_{t-\tau}, \tag{172}$$

where B_τ and A_τ are the matrices of coefficients (B_τ is a square matrix), whereas y_t and x_t are the two vectors with components respectively given by the y_{jt} and the x_{it}. We note moreover that h is the maximum lag appearing in the equations system.

It is no loss of generality to write Equations (171) without additive constant terms because, if needed, we can define, the last exogenous variable x_{mt} as being the "dummy variable" permanently equal to 1 and we can fix $a_{km\tau} = 0$ for all positive τ: the coefficient a_{kmo} in the k-th equation then plays the role of a constant term.

When reading the compact vector equation (172) we must be aware of the fact that there are two kinds of elements in the matrices B_τ and A_τ as specified by the model: those elements which are assumed known and therefore appear as numbers already in the specification, and those which appear as parameters, to be estimated for applications of the model[165]. Typically all coefficients $b_{kj\tau}$ and $a_{ki\tau}$ of an accounting identity k are known (although the rate of depreciation δ in the identity (151) may be taken as a parameter, or alternatively (151) may be taken as a behavioural equation, in the wide sense given to the latter expression here). A behavioural equation k typically contains parameters, but also known coefficients; usually many more: for instance, we know that, for all values of the index τ, $a_{ki\tau}$ is equal to zero if the exogenous variable i did not appear in the specification of equation k before this equation was embedded in the formal representation (171) of the system; we also know $a_{ki\tau} = 0$ if variable i did not appear in equation k with lag τ. Moreover for definiteness, each equation in (171) must contain at least one known non-zero coefficient:

[165] For simplicity in this exposition we rule out known constraints on the parameters, such as the fact that two given parameters sum up to 1, or such as an inequality on the value of a parameter.

without changing in any way the meaning of an equation and what would be known about it, its parameters could be all multiplied by the same arbitrary non-zero number if all coefficients of this equation were either known to be zero or entirely unknown parameters.

Finally we normally mean the system (171), or equivalently (172), to provide a full determination of its endogenous variables from its predetermined variables: precisely, we mean that the n values of the y_{jt} (for $j = 1, 2, \dots, n$) are determined through (171) from the $(n+m)h + m$ values of the $x_{i,t-\tau}$ (for $i = 1, 2, \dots, m$; $\tau \geqslant 0$) and $y_{j,t-\tau}$ (for $j = 1, 2, \dots, n$; $\tau > 0$). This is the case if and only if the square matrix B_0 has rank n. We assume that such is the case, so that we can derive from (172):

$$y_t = \sum_{\tau=0}^{h} B_0^{-1} A_\tau x_{t-\tau} - \sum_{\tau=1}^{h} B_0^{-1} B_\tau y_{t-\tau}. \tag{173}$$

This vector equation shows how the values of the y_{jt} can be computed as soon as the values of the matrices $B_0^{-1} A_\tau$ and $B_0^{-1} B_\tau$ have been estimated.

The methodology of structural models speaks of "*the reduced form*" to designate (173) in contrast to "*the structural form*" (172). Individual equations of (173) are said to be "reduced equations", in contrast to the structural equations gathered in (172). We easily understand why structural equations are often said to have more autonomy than reduced equations: any of the latter equations does not make sense without reference to the full system (171) from which it was derived. We also understand that the elements of matrices $B_0^{-1} A_\tau$ and $B_0^{-1} B_\tau$ of the reduced form are complex expressions of the parameters appearing in B_0 and in the corresponding matrix A_τ or B_τ.

3.3. Estimation of macroeconometric models

3.3.1. *The data base*

How are the parameters to be estimated? The answer was given unambiguously by promoters of structural macroeconometric models: estimation will be given by a fit of the model to a set of time-series collected from the statistics of the country concerned. More precisely the data base must be made of $m + n$ time series, one for each variable x_{it} or y_{jt}, these series covering a past period from an initial date t_0 to a terminal date $t_0 + T$, the last one for which data exist. The series is annual or quarterly, depending on the length of the unit period specified in the model. Precise definitions, in terms of the national accounts and statistics, have been chosen for each variable. Comparability of the data through time is insured, as well as consistency with all accounting identities of the model.

Such a data base is adequate but narrow. It is adequate because the definitions of the variables make sense for users of the model, so that, for instance, macroeconomic forecasts can be expressed directly for future values of the set of endogenous variables (or for its most interesting subset) from independently forecast values of exogenous variables. In other words, the responsibility for choosing the adequate economic language has been given to the builders of the model, who had to be aware not only of macroeconomic theory but also of both users' needs and availability of data to be provided by statisticians and national accountants.

But the data base made of these $m + n$ time series is narrow, in the sense that it does not encompass all the empirical knowledge that is relevant with respect to estimating the values of the parameters. In the first place, it ignores similar data bases which exist for other neighbouring countries, notwithstanding the likelihood that about the same behaviours there prevail. It also ignores all the microeconomic data bases available in the country, whereas the content of these bases may be particularly informative with respect to some of the modules assembled in the macroeconomic model.

3.3.2. *The stochastic model*

Given the data base the econometrician cannot expect the fit on system (171) to be perfect. He or she must recognize the existence of discrepancies between the values taken respectively by the right- and left-hand members of each behavioural equation k in (171). This might be in part due to errors of measurement in the values recorded for variables x_{it} and y_{jt}. Notwithstanding the existence of such errors, macroeconomists think they can usually be neglected because a much more important effect is likely to come from unspecified causes affecting the behaviour represented by equation k. Indeed, we had an idea of the large number of factors that may play a part in the macroeconomic observation of behaviours when we studied each type of behaviour or adjustment law, many of these factors being not even observed. Within models of the whole economy only the main factors are explicitly introduced.

A complete formalization of the phenomenon represented by an equation k requires recognition of the effect of all omitted causes. Within linear models it is natural to introduce this effect in the form of an additive term ε_{kt}. It is then understood that ε_{kt} is not observed and can be considered as random. It is also common to assume that the mathematical expectation of ε_{kt} is equal to zero, which is an innocuous assumption when equation k contains as parameter an additive constant. System (171) is then replaced by:

$$\sum_{j=1}^{n}\sum_{\tau=0}^{h} b_{kj\tau} y_{j,t-\tau} = \sum_{i=1}^{m}\sum_{\tau=0}^{h} a_{ki\tau} x_{i,t-\tau} + \varepsilon_{kt}, \quad k = 1, 2, \ldots, n. \tag{174}$$

We speak of a system with "errors in equations". (In the general formalization of the system by (174) we know that $\varepsilon_{kt} = 0$ if equation k is an accounting identity.)

Econometric theory devoted a considerable attention to the estimation of linear systems of this form. Econometricians working on the structural models here discussed apply more or less elaborate techniques, the properties of which have been carefully studied under specified stochastic assumptions. A presentation of these techniques and of their properties would much exceed the confines of this book[166]. We must, however, briefly consider the identification problem. We must also be aware of two rather general conclusions which follow from experience with national macroeconometric models.

Let us remark in passing that we already saw in Chapter 8 two particular cases of estimation of models having form (174) but not claiming to cover the whole economy. In Part 6 we discussed the econometrics of the price and wage equations, which corresponds to the treatment of a case with two endogenous variables ($n = 2$), an index of the price level and an index of wage rates. More instructive now is Section 5.5, in which, considering how to empirically assess the effectiveness of monetary policy, we proceeded to a first examination of vector-autoregressions, the VARs. The equations then involved only few endogenous variables, but the main distinctive feature was the absence of exogenous variables: the double sums in the right-hand member of Equations (174) had disappeared. (Moreover, there was no accounting identity.) This is a quite different case from what we see in structural macroeconometric models, which usually contain many exogenous variables.

3.3.3. *The identification problem*

However, the presentation in Section 5.5 of Chapter 8 is now interesting because it devoted attention to the identification problem. We then realized that

[166] The reader may find an introduction to the relevant econometric theory in Part 5 of E. Malinvaud, *Statistical Methods of Econometrics* (North-Holland, 3rd ed., 1980).

the set of the $n^2(h+1)$ coefficients $b_{kj\tau}$ was indeterminate if we knew nothing about the probability distribution of the vector ε_t and about the values taken by the coefficients, because premultiplying by an arbitrary regular square matrix C all matrices B_τ and the vector ε_t, while keeping unchanged the values of the vectors $y_{t-\tau}$, would not perturb the equality between the right- and left-hand members of the equations. In other words, observing the time series of the vector y_t could not suffice for unambiguously estimating the matrices B_τ. Indeterminacy, i.e. lack of identifiability of the matrices B_τ from the observation of the time series $\{y_t\}$, would subsist if a normalization rule was introduced for each equation, such as $b_{jj0} = 1$, and if general properties of the stochastic process $\{\varepsilon_t\}$ were assumed, such as the hypothesis that the vectors ε_t were independently and identically distributed with a zero mathematical expectation.

We then noted that various solutions to the identification problem had been used, all of them amounting to adding further restrictions of the system. We paid particular attention to the case in which the system is assumed to be "recursive" for a particular ordering $j = 1, 2, \dots, n$, recursiveness meant that the matrix of the second order moments $E\varepsilon_t\varepsilon_t'$ was diagonal and the matrix B_0 was triangular:

$$b_{kj0} = 0 \quad \text{for all } j > k \tag{175}$$

(all the coefficients placed above the first diagonal of B_0 were known or assumed to be equal to zero).

When now attention turns to structural macroeconometric models, we are aware that lack of identifiability might be a problem. A priori the situation may seem to be less acute, because we saw that many coefficients were assumed to be equal to zero in the specification of the model. Indeed, in a number of cases recursiveness was shown to hold for the system (174), the matrix B_0 being triangular and the matrix $E\varepsilon_t\varepsilon_t'$ being diagonal. System (174) is then identifiable for the same reasons which make VAR recursive systems identifiable.

Another simple case of identifiability occurs in structural models when each equation k contains a specific exogenous variable, i.e. a variable x_{it} which does not occur in any other equation. If such specific exogenous variables are ordered before the other exogenous variables and in the same order as the endogenous variables to which they correspond, the property means:

$$a_{ki0} = 0 \quad \text{for all } i \neq k. \tag{176}$$

When the property is known to hold, the structural model is again identifiable.

But realization of property (175), together with diagonality of $E\varepsilon_t\varepsilon_t'$, or of property (176) is rather exceptional in actual structural models, in the same way as symmetries are. Identification problems occur in some cases. More importantly in macroeconomic practice, identifiability may be difficult to assess off hand. This is why identifiability criteria have been developed.

3.3.4. *Overidentification*

Actually with macroeconometric models we are usually in a situation that may be said opposite to that of VAR models: not lack of identifiability but rather overidentification, not too few but "too many" prior restrictions meaning that some variables should not occur in some equations (we shall come back to this opposition in Section 4.4; for the time being we take it as a fact, which was not disputed by practitioners before the end of the 1970s). Too many with respect to what? To easiness in estimation.

In order to understand the point, let us go back to the correspondence between the structural form and the reduced form. Remembering the end of Section 3.2 we know that, given any values of the matrices B_τ and A_τ such that B_0 is not singular, we may associate to (174) a reduced form:

$$y_t = \sum_{\tau=1}^{h} Q_\tau y_{t-\tau} + \sum_{\tau=0}^{h} P_\tau x_{t-\tau} + \eta_t, \tag{177}$$

where the new matrices and the new random vector are given by:

$$P_\tau = B_0^{-1} A_\tau, \qquad Q_\tau = -B_0^{-1} B_\tau, \qquad \eta_t = B_0^{-1} \varepsilon_t. \tag{178}$$

Equation k in (177) gives the value of the endogenous variable y_{kt} as a function of the error η_{kt} and of the predetermined variables: the exogenous $x_{i,t-\tau}$ for $\tau = 0,\ldots,h$ and the lagged endogenous $y_{j,t-\tau}$ for $\tau = 1,\ldots,h$. The vector equation (177) looks convenient for estimation.

It would indeed be so if the $2h+1$ matrices Q_τ and P_τ could take any values within their respective formats $(n \times n)$ or $(n \times m)$. If so, we could think of fitting each equation k in (177) by a linear regression of y_{kt} on all predetermined variables. This would give estimates Q_τ^* and P_τ^*, from which corresponding values B_τ^* and A_τ^* would be found by solution of the matrix equations in (178). (We assume for simplicity, up to the end of this section, that moments of ε_t of higher order than 1 are unrestricted.)

With respect to this approach it is easy to prove that a multiplicity of solutions of the set of matrix equations exists when the model (174) is not identifiable and that no solution exists, except by chance, when actually the matrices

Q_τ and P_τ are subject to restrictions, which have been ignored in the regressions but result from the definition of the matrices by (178). Heuristically, we can say that the first case occurs when there are too few prior restrictions on the coefficients of the structural form and that the last case occurs when there are so many such restrictions that they restrict, through (178), the values that the matrices appearing in the reduced form can take. When this last case occurs, we say that the structural model is overidentified. Indeed, there is also an intermediate case of just identified models, in which the indirect regression approach described above perfectly applies and leads to a unique estimation of the structural form.

Overidentification is the common case in macroeconometric models. A number of estimation techniques have been defined in econometric theory for this case, and their properties have been studied. We shall not go into this theory, but simply note that most of its practical recommendations are based on asymptotic properties, that is, on properties that have been proved to hold at the limit, under fairly general sets of assumptions, if by hypothesis the number of periods T in the data base would moreover increase indefinitely.

3.3.5. *Regressions on structural equations*

When he worked in the 1930s on the first national macroeconometric models to have ever been specified and estimated, Jan Tinbergen ignored this econometric theory, which did not exist at the time. He simply fitted independent regressions on each one of the structural equations making his system, i.e. on each behavioural equation k of (171), taking account of the restrictions applying to this particular equation. In other words, he took into account that some of the coefficients $b_{kj\tau}$ and $a_{ki\tau}$ of our general formalization did not appear in his equation k; he had chosen a normalization, let us say $b_{kk0} = 1$; he so regressed y_{kt} on all the other variables appearing in his equation.

Such independent fits of structural equations were made by others since then, even after the econometric theory of estimation in structural systems had been substantially developed. The first tenet to have been established by this theory was, however, that such independent fits were not appropriate. The tenet provided the main impetus for the definition and study of the alternative techniques to which we just referred at the end of the last subsection.

Such being the case, it is ironical to now have to record that experience of several decades with national macroeconometric theory lead to a more guarded assessment: Tinbergen's practice of independent fits of structural equations was risky and still is when used by others; but its results turn out to be in many cases about as good or as bad as the ones reached by application of supposedly better-founded estimation techniques. How can such an assessment be explained?

The explanation turns out to illustrate the simple fact that asymptotic properties derived in mathematical statistics do not have universal significance. As we saw a moment ago, when studying estimates in structural models econometric theory long focused on the limit properties, which would hold if the length T of the series in the data base would increase indefinitely. Specialists knew that this length was not only finite but even rather short, considering the size of the residuals remaining in equations after estimation of the coefficients. But specialists were long unable to provide useful information about the "small-sample properties" of alternative estimators under consideration. Indeed, the probability distributions of these estimators for a given T were complex to characterize and still more difficult to compare. In contrast, asymptotic theory could be worked out and provided clear results.

The first result stated that, in general, estimators provided by independent regressions on structural equations were asymptotically biased: they were converging to values that differed from true values, this under natural hypotheses about the error vector ε_t, in particular that it was uncorrelated with the vectors of predetermined variables ($x_{t-\tau}$ for any $\tau \geqslant 0$, $y_{t-\tau}$ for any $\tau > 0$). In contrast, the alternative estimators proposed by econometric theory were "consistent", i.e. asymptotically unbiased: they were proved to converge for each coefficient to the true value of this coefficient in the model that was assumed to have generated the data and to have been correctly recognized by the econometrician.

Also important were results providing users with precise methods of inference: estimators were proved to have asymptotic Gaussian distributions with variance-covariance matrices for which formulas were given. Under fairly general assumptions these covariance matrices had the size T of the sample in their denominator. They could be estimated thanks to formulas which were proved to be asymptotically valid. Thus, users could also perform tests of hypotheses about the values of some interesting coefficients.

With respect to the comparison between Tinbergen's practice and the asymptotically more valid methodologies that were proposed, it was already relevant to draw a first conclusion: most often the covariance matrices of the consistent estimates were larger than those resulting from independent regressions on structural equations. Let us follow current practice and call the latter OLS estimates ("ordinary least squares"). Thus, for a finite sample size T, OLS estimates were biased but less dispersed than were consistent estimates. Since the bias was not decreasing to zero when T was increasing, a sufficiently large T

would certainly make the consistent estimates preferable to OLS. But, since T was usually small, the case was less clear; in particular, the mean squared estimation error of a coefficient (the square of the bias *plus* the variance) could be smaller with OLS.

This remark was confirmed when research in econometric theory aimed at finding closer approximations to small sample properties than those given (to the order $T^{1/2}$) by the initial asymptotic theory. It was found in closer approximations (to the order T^{-1}, for instance) that distributions of consistent estimates tended to have fatter tails than OLS had[167]. Monte Carlo simulations on small structural models led to similar conclusions, suggesting that in practice OLS was not as bad, relatively to other available estimators, as had been earlier thought.

Other results also contributed to a prudent rehabilitation of OLS, namely those showing that, in practice and considering the substantial standard deviations of the estimates, most results reached with OLS were not so different from those reached with seemingly better-founded consistent methods[168]. For applications in forecasting and policy analysis, the main problem was rather the low degree of accuracy of the results, whatever the method from which they came. This is indeed the second general conclusion to be drawn from several decades of experience with structural macroeconometric models. We shall have more to say about this conclusion in Section 3.5 and later in Section 4.3.

Before leaving problems of estimation, let us be clear on the implications to be drawn from the bird's-eye view given in this section. The careful analysis provided by econometric theory should not be disparaged and presented as useless for macroeconomists. On the contrary, without it macroeconomists would not really know what they could learn from macroeconomic time series concerning complex societies. Moreover, we must not forget that OLS can be very misleading in some actual cases and that only familiarity with econometric theory can permit us to detect where these cases stand.

[167] For precise studies of such properties see in particular T. Anderson and T. Sawa, Distributions of estimates of coefficients of a single equation in a simultaneous system and their asymptotic expansions, *Econometrica* (July 1973); T. Anderson, An asymptotic expansion of the limited information maximum likelihood estimate of a coefficient in a simultaneous equation system, *Journal of the American Statistical Association* (June 1971); T. Anderson, K. Morimune and T. Sawa, The numerical values of some key parameters in econometric models, *Journal of Econometrics* (February 1983).

[168] See C. Christ, The Cowles Commission's contributions to econometrics at Chicago, 1939–1955, *Journal of Economic Literature* (March 1994), particularly Section 7.g.

3.4. Econometric models as forecasting instruments

3.4.1. *The theory and the practice*

In principle, there is no difficulty in explaining how a macroeconometric model can be used in forecasting. We already mentioned it earlier. According to the model, some exogenous variables determine endogenous variables. The logical approach has two natural steps: first, to forecast the values of the exogenous variables by any relevant method, second, using these forecast values, to solve the system of equations which constitute the model so as to obtain the values of the endogenous variables for the same future period.

In the case of the DMM model, for example, we have to forecast government demand G and the monetary base $\overline{M}$, two variables which characterize economic policy, about which some public declarations are usually available. We also have to consider trends in international trade and deduce from this a forecast for Ex. Finally, we have to determine what the labour supply $\overline{L}$ will be. Solution of the model gives the volumes of production, consumption, investment and imports, together with the levels of prices, unemployment, the trade deficit or surplus, the money supply, etc.

In the same way, the model can in principle be used to study the effects of various economic policy proposals. Indeed, we can calculate as many forecasts as there are proposed policy variants, and we can compare them with each other. But such a policy analysis raises new questions which will be studied in Section 3.6. At this point we limit attention to pure forecasts, except for marginal comments in passing.

Direct use of a model immediately calls for three remarks of a formal nature. In the first place, to say that endogenous variables are determined by exogenous variables is an abuse of language, since behavioural equations are not exact, as we recognized in the foregoing section: behavioural equations claim to be valid on average but not in each period. In other words, we have to add to each behavioural equation an additional random term showing those effects which are either not identified or else not taken into account. This term is assumed to have zero mathematical expectation, since without it the relationship is considered as "valid on average". Neglecting, in the forecast of endogenous variables, the presence of all random terms is accepting errors in the forecasts, but this is unavoidable. The choice of a particular forecasting method, about which we shall be more explicit in Section 3.6, does not remove the difficulty.

In the second place, although the econometric model contains as many equations as endogenous variables, we might fear difficulties when solving the system of equations. Will we always find a solution? Might we not, perhaps, find several? These questions arise all the more since almost any macroeconometric model includes non-linear equations. This is easily explained: values are included along with quantities and prices, ratios and rates of change appear together with the magnitudes concerned, some behavioural equations are directly posed as non-linear. In practice, these features are only serious in that they complicate the calculations, but the system of equations is always found to contain one, and only one, solution in the region which is a priori known to contain the future vector of the endogenous variables.

In the third place, alongside the endogenous variables for the period on which the forecast bears, appear not only truly exogenous variables but also "lagged values" of endogenous variables, that is, values referring back to former periods: the DMM model includes such lagged values in many equations. Hence, we can interpret the model in two different ways. It can be conceived as determining endogenous variables from "predetermined variables", which include, by definition, exogenous variables and lagged values of endogenous variables. A model interpreted in this direct way can only be useful in forecasting the following period, once the values of the lagged endogenous variables are known. But we can also interpret the model as defining the dynamic system which determines the paths of the endogenous variables from the paths of the exogenous variables and from the initial values (that is, known former values) of endogenous variables. Thus, forecasting can be made step by step and extended over several future periods.

But these three formal remarks are only preliminary when we have to describe the complexity in the actual practice of those responsible for making macroeconomic forecasts and for comparing the future effects of different variants of economic policy. These economists never rely only on their macroeconometric model.

In a broad sense, their approach is fundamentally similar to that which is imposed by the model. Indeed, they have to first forecast exogenous changes. This means taking demographic trends into account, studying perspectives for the international environment, specifying government policy in macroeconomic terms, and so on. They then have to derive the consequences, which result also from agents' behaviour and from the requirements of accounting consistency. But the approach is usually applied less rigidly than reliance on a standing model would imply. Sometimes forecasters dispense altogether with econometric models, sometimes they reconsider and "adjust" the results given by their model[169].

[169] A good account of work carried out in practice by economists who had to make forecasts and prepare French economic policy in the 1960s was presented by B. Billaudot and F. Lagrange,

The various reasons for making such adjustments may be understood. First of all, the user may not be fully confident in his model. Building any econometric model takes time and means that a level of aggregation and many other simplifications are fixed once and for long. Yet a critical understanding of the economy is the result of continual analysis, at different degrees of aggregation and based upon always new investigations. The model may well be somewhat outmoded by the progress of recent research. Owners of the model may even have observed systematic departure of observations from what would be implied by the model. A consistent tendency for an equation to overpredict or underpredict in the recent past may be interpreted as a first sign of specification error, to which a quick solution is to assume that the error will continue into the future, and so to add back the current residual, or an average of recent residuals, as an adjustment in the forecast period.

Above all, the model may be inadequate for the special circumstances in which forecasts have to be made. In order to understand the point, let us simply recall what French economists in charge of macroeconomic forecasting had to consider, even simply when Keynesian analysis was thought to be essentially reliable: the consequences of the international tension following the English and French closing of the Suez canal in 1956, the return of about one million persons who were repatriated to France from North Africa in 1962, the student and worker protest in 1968, the increase in the price of oil in 1973, the removal of price controls in 1978. These examples are typical of the many circumstances which are not envisaged at the time at which a model is constructed, and which later may raise doubts as to how adequate the model is when faced with the circumstances in question (in the examples above, were the models in use in 1956, 1962, 1968, 1973 or 1978 still adequate for the forecasts made for 1957, 1963, 1969, 1974 or 1979?)[170]. Someone who is badly informed al-

Prévisions et dévaluation de 1969, *Statistiques et études financières*, quarterly series, No. 8 (1972). Also recommended is M. Surrey, *The Analysis and Forecasting of the British Economy* (Cambridge University Press, 1971), which shows the role played in practice during the 1960s by the use of formal analysis.

[170] S. McNees, The role of judgment in macroeconomic forecasting accuracy, *International Journal of Forecasting* **6** (1990) 287–299, illustrates the same point from the American experience: "The economy is more or less continually buffeted by a variety of events if not entirely unique, a least without close historical precedent. Most recently, it has been drought, earthquake, freeze, and abnormal warming. A decade ago, rather than natural disasters it was strikes, oil shocks, and the phasing in and out of wage, price, and credit controls. At other times, institutional and regulatory changes have led to changes in the very definition of macroeconomic concepts. Such events have implications for the economic future, implications not well captured by standard models which, of necessity, describe the "normal", past behavior of the economy. The evidence presented here broadly confirms the conclusion that individuals adjust their models to compensate in part for their models' deficiencies, thereby improving the accuracy of their forecasts" (p. 298).

ways exaggerates the importance of the breaks that such events cause; in fact, they do not make previously constructed models fundamentally inadequate; but they explain why the user of the model may have to revise somewhat numerical results after a careful study of some relevant aspects which builders of the model ignored.

In addition, the user may have much richer information than that exploited by the model. Thus, he or she may have direct data on firms' intentions as to their future investments (in most countries statistical enquiries are made regularly to determine these); at times these data may lead to an aggregate evaluation which is quite different from the one resulting from the assumed behavioural equation. In such cases, the user contemplates the possible explanations for such a discrepancy; this may lead him or her to use the direct information in preference to the verdict of the behavioural equation which was meant to hold on average. In such a case the user may simply introduce an appropriate non-zero error into the corresponding equation.

To these various reasons for the introduction of non-zero errors into the model, we must add another, which leads to a more automatic correction. Observation often shows that deviations with respect to average behavioural equations are serially correlated: a positive deviation for one period is often followed by a positive deviation in the next period (the case of alternation between positive and negative deviations can also happen for some behavioural equations). An econometrician who has to construct a model must study whether the random errors recognized in the model are serially correlated. If so, he or she must make recommendations to users on how to correct for this correlation in the computation of forecasts; this will normally be by the addition of non-zero errors, whose values will depend on the positive or negative residuals noticed in the course of the most recent periods.

Using a model therefore does not dispense, in practice, with analyses which are less mechanical and more specific to the current situation or to problems which may arise when particular economic policy measures are contemplated. Likewise, when doing without a model, we still have to respect accounting consistency and take into account what studies over the past have taught us about behaviour. The main advantage of a model for a team involved in making economic forecasts is to provide a rigorous instrument, which is moreover as transparent as possible. It is a useful anchorage when the team wants to make good use of the progressive accumulation of knowledge about the economy.

3.4.2. *Accuracy of macroeconomic forecasts*

What has been the success of forecasts obtained from macroeconometric models? This is a natural question to ask and the answer is not so different from the one to the more general question: how well did economic forecasts perform? Such questions concern the outcome of practices developed by people in charge of forecasting. Evaluation of econometric models, which we shall discuss in the next section is a different subject since models are not used only in forecasting and forecasters do not use their model only.

There is a literature on forecast errors, ranging from descriptive accounts in specific periods, especially at business cycle turning points, to statistical analysis of successive forecasts over a number of quarters or years. Here for broad answers to the questions raised, we naturally adopt the latter approach. The landmark from which a detailed survey would begin is the book of H. Theil[171], which contains almost 400 pages to present and apply a methodology for the evaluation of forecast accuracy. His results about the accuracy actually reached at the time may be found too old to be now relevant, but the methodology did not change so much over the last forty years.

Let us here consider an observed time series y_t for a sequence of periods t. Let y_t^f be the forecast of y_t made by the economist or the agency f in period $t-l$; for simplicity the forecast lead time l is assumed to be the same for all forecasts studied and is not introduced in the notation. Let the forecast errors e_t be known for a sequence of T periods, with of course:

$$e_t = y_t - y_t^f, \qquad t = 1, 2, \dots, T. \tag{179}$$

The statistical analysis bears on the properties of the time series $\{e_t\}$ and on its relations to other data, particularly the time series $\{y_t\}$.

The natural first idea is to define a measure of the overall importance of the errors. Sometimes the mean absolute error (MAE), i.e. the mean of the T absolute value $|e_t|$, was chosen. But it is more common to consider the root mean square error (RMSE) defined by:

$$(RMSE)^2 = \frac{1}{T}\sum_{t=1}^{T} e_t^2. \tag{180}$$

This may be justified as a rough measure of the cost that forecasting errors imply. Many other characteristics may also be considered; for instance, a simple average of e_t may be taken as a measure of the bias in the forecast[172].

[171] H. Theil, *Economic Forecasts and Policy* (North-Holland, Amsterdam, 1958).

[172] For other measures and a much fuller discussion see K. Wallis, Large-scale macroeconometric modeling, in: H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics – Macroeconomics* (Backwell, Oxford, 1995).

Results of such statistical measures do not suffice for an evaluation of the series of forecasts in question. They must be compared with the results that other methods of forecasting would have given. In macroeconometrics there is a long tradition of using so-called "naïve" forecasts constructed without reference to any economic theory and usually based on past values of the variable y_t. Gradually these have become less naïve. In the old days the alternative forecast y_t^g was simply "no change": y_{t-l}; soon it became "same change", something like: $2y_{t-l} - y_{t-2l}$. Since the beginning of the 1970s the alternative forecast was often supposed to be given by a time-series statistician extrapolating $\{y_t\}$ as if it was the realization of some stochastic process: an autoregressive process *plus* a trend, or even an ARIMA process. But this increased sophistication of the benchmark is not the only reason why confidence in macroeconomic forecasts decreased twenty years ago after a period during which they had seemed to work fairly well.

Speaking at the end of 1977, V. Zarnowitz[173] could start with a reminder that "the period of transition from the war economy witnessed the largest errors in the US GNP forecasts": declines were forecast in 1947 and again 1948, whereas a very fast growth was recorded. But during the 24 years from 1953 to 1976, the mean absolute error for annual mean predictions of the percentage change in nominal GNP by private forecasters was 1.2 per cent, whereas the mean absolute error for a naïve extrapolation was 2.3 per cent (next year's percentage change in GNP will be the same as the average percentage change in the four previous years). Similarly during the 15 years from 1962 to 1976, the mean absolute errors for annual predictions of change in real GNP and the GNP price deflator in the Economic Report of the President were respectively 1.1 and 1.0, as against 2.6 and 1.3 in a naïve extrapolation (next year's percentage change will be the same as that of the previous year). About the same results were obtained with forecasts made during the same period by a group using a macroeconometric model in the University of Michigan.

But in the late 1970s, as quarterly models were commonly used and as extrapolation by time-series statisticians were becoming more and more sophisticated, the comparison became less favourable to forecasts derived from structural models. The common wisdom which then emerged was expressed by L. Klein[174]: "There is evidence, however, that time series models do perform about as well in forecasting as the adjusted macroeconomic model in very short horizons, say up to 3 months or possibly up to 6 months".

[173] V. Zarnowitz, On the accuracy and properties of recent macroeconomic forecasts, *American Economic Review* (May 1978).

[174] L. Klein, The importance of the forecast, *Journal of Forecasting* (January–March 1984).

During the first half of the 1980s the diagnosis became actually still less favourable to adjusted forecasts derived from structural models. Not only errors of those forecasts tended to increase, but also the VAR movement promoting its approach to macroeconomic forecasting could register fairly good performance, particularly for what was called the Bayesian VAR (BVAR). With respect to the extrapolations proposed by time series statisticians two improvements were introduced: (i) a few series were processed jointly instead of independently of one another (this was the main idea of the VAR methodology, as we saw in Section 5.5 of Chapter 8); (ii) extrapolations were corrected so as to be somewhat attracted toward *a priori* chosen central values (this was, transposed to a multivariate dynamic context, the same idea as that inspiring "ridge regressions"[175], which can be justified by the Bayesian approach to statistical estimation or otherwise). Be that as it may, analysis of the macroeconomic forecasts made in the US for the quarters 1980.2 to 1985.1 gave disturbing results which were presented and commented independently by S. McNees and R. Litterman[176].

Studying precisely the quarterly forecasts made by eleven US prominent forecasters with a forecasting lead time running up to eight quarters, including the BVAR forecasts issued by R. Litterman, S. McNees first found the surprising result that for many macroeconomic variables the RMSE tended to decrease when the lead time was increasing (exceptions were the rate of inflation, the interest rate and the unemployment rate). Second, overall the accuracies reached be different forecasters did not permit a simple ranking because the ranks varied from one variable to another. But, third, forecasts made from structural models appeared in many cases worse that those following from the BVAR: the BVAR forecasts were generally the most accurate or among the most accurate for real GNP, the unemployment rate and investment; they were the least accurate, especially at long horizon, for the GNP deflator. McNees argued that probably the forecasts of all forecasters except the BVAR took too seriously in their adjustments the announces made in 1981 by the newly elected President about what he would achieve. He also speculated as follows: "Just as conventional macroeconometric models may have been oversold in the 1960s and early 1970s, leading to disappointment and rejection in the late 1970s and 1980s, there is some danger that the VAR approach to modeling and

[175] On ridge regressions see, for instance, E. Malinvaud, *Statistical Methods of Econometrics*, 3rd ed. (North-Holland, Amsterdam, 1980).

[176] S. McNees, Forecasting accuracy of alternative techniques: a comparison of US macroeconomic forecasts, R. Litterman, Forecasting with Bayesian vector autoregression: five years of experience, both in *Journal of Business and Economic Statistics* (January 1986).

forecasting may now be oversold as a superior substitute for the more traditional approaches".

Indeed, in an extended comparison over the whole of the 1980s, McNees (1990, op. cit.) found a change in the relative performance of the BVAR between the first and the second half of the 1980s; this appeared most clearly about investment, the BVAR being the most accurate in the early 1980s but the least accurate in the late 1980s, in both cases by sizeable margins; the relatively poor performance of the BVAR forecast of inflation deterioted further.

The historical account just reported is limited to the US experience. But the common wisdom is to think that similarly serious and extensive studies of the European experience would record roughly similar conclusions if such studies existed. In the next section we shall consider extra material which, concerning the UK but being differently focused, will be consistent with this common wisdom.

3.4.3. *Sources of forecast errors*

Since in practice macroeconometric models are not alone to be used in the production of forecasts, we should like to know what is their contribution to the accuracy of forecasts. One approach, which will be now examined, considers a decomposition of forecast errors so as to isolate the part due to the model. Alternative approaches will be examined in the next section when we shall deal with the evaluation of models.

The method for the decomposition has been exposed and applied by K. Wallis and J. Whitley[177]. It focuses on two reasons for errors, besides what may be due to the model: the forecaster must, independently of his or her model, project values of the exogenous variables, values which will turn out to be more or less erroneous; the forecaster also has the liberty, even the responsibility, to exert his or her judgement and to see whether particular adjustments to the equations of the model are not advisable (with models of some size it would be exceptional that not a single adjustment be advisable), but again more or less noticeable errors of judgement may occur and will at times do.

For their decomposition Wallis and Whitley use a formal notation, which we may repeat. Let x be the vector of exogenous variables, or more precisely the value that the vector actually turns out to have. Let $\hat{x}$ be the value which was *ex ante* forecast for this vector. Let similarly a be an index defining which judgemental adjustments were *ex ante* brought to the model, with $a = 0$ meaning no adjustment. Let $\hat{y} = f(\hat{x}, a)$ be the *ex ante* forecast of an endogenous

[177] K. Wallis and J. Whitley, Sources of errors in forecasts and expectations: UK economic models 1984–8, *Journal of Forecasting* **10** (1991) 231–253.

variable, whose *ex post* value actually is y. So the error in the forecast of this variable is naturally denoted as $y - f(\hat{x}, a)$, and the part imputed to the model is $y - f(x, 0)$. Wallis and Whitley then use the following decomposition:

	model error	$y - f(x, 0)$,
less	contribution of adjustment	$-[f(x, a) - f(x, 0)]$,
plus	contribution of exogenous variable errors	$+[f(x, a) - f(\hat{x}, a)]$,
equals	error in published forecast	$= y - f(\hat{x}, a)$.

To avoid confusion with what will come later in our discussion, we remark in passing that $f(x, a)$ is called the "*ex post forecast*": the value that would have been forecast for the endogenous variable if exogenous variables had been perfectly forecast. Wallis and Whitley use the expression "hands-off *ex-post* forecast" to designate $f(x, 0)$.

Clearly, an intimate knowledge of the work of the forecaster is needed in order to compute the decomposition. The full vector $\hat{x}$ of *ex ante* forecasts of the exogenous variables is not commonly published; the version of the model that was used often differs from the one earlier published; judgemental adjustments are only partially reported, if at all. But for reasons which will appear in the next sections, Wallis and Whitley had access to enough direct information to be able to compute and publish decompositions for forecasts made by four teams using each a quarterly macroeconometric model of the UK. The forecasts so decomposed were made in 1983, 84, 85 and 86 at the horizon of one year, and even two years for three of the teams; they concerned two endogenous variables, the level of Gross Domestic Product and the inflation rate; there was thus 31 decompositions for each of the two variables.

The most noticeable results show that published forecasts are superior to the hands-off *ex post* model forecasts; this fact points to the beneficial influence of adjustments. In 20 cases out of 31 for GDP and in 24 cases for inflation $|y - f(\hat{x}, a)|$ is smaller than $|y - f(x, 0)|$. The root mean square errors are respectively 2.8% and 4.6% for GDP, 0.9 and 4.4 percentage points for inflation. The decomposition indeed shows that the contribution of adjustment has most often an opposite sign to the model error (20 cases for GDP, 25 cases for inflation).

In this application, the contribution of errors in the forecast of exogenous variables does not play the expected role: only in 11 cases for GDP and in 13 for inflation would a better accuracy have resulted from the substitution of actual values of exogenous variables to forecast values. Surprisingly, the authors report that this result is not peculiar to the UK models; they quote

Klein writing about the "well established fact that errors associated with *ex ante* forecasts tend to be much smaller than those of *ex post* extrapolations". They see in this fact an effect of model misspecification.

Working on US model-based forecasts S. McNees (1990, op. cit.) was in a less privileged position than Wallis and his associate: he had no access to the model themselves. But, for four prominent forecasters which published appropriate data, he could compare the main (publicized) forecasts resulting from their adjusted models to the forecasts which were provided by the same models when operated mechanically without adjustment. He then found that, with some significant exceptions, judgemental adjustment improved accuracy more often than not. He also investigated the question whether these forecasters were optimally weighting the judgemental evidence leading to their adjustments; he then found a weak tendency for modellers to overadjust their models (psychologists observed, more generally, that in their intuitive predictions people were giving too much weight to specific circumstances capturing their attention).

3.5. Analysis of macroeconometric models

Even though they cannot claim to be precise, full macroeconometric models do offer quantitative information which merits serious examination. In dealing with the short-term equilibrium, with the short-run dynamics of output and inflation or with business cycles earlier in this chapter, we looked at a large number of models which were at times purely qualitative, at times econometric but partial. Leaving aside for the moment the VAR models promoted in the 1980s and the calibrated models advocated more recently by the RBC movement, we find in the macroeconometric models now discussed an instrument available for a wide range of quantitative applications in which all the main economic interrelations matter. This instrument claims to provide, in its country or area covered, reliable measures of the effects of changes in its exogenous variables on its endogenous variables. Examining these measures and their reliability is certainly worth-while. Such is the purpose of this section.

Two approaches will be applied. We shall, first, aim mainly at taking stock of the measures obtained, that is, of the resulting quantitative relations between exogenous and endogenous variables. The multipliers will characterize the strength of the effects emanating from changes in the main exogenous variables. The time profile of the effects will reveal their lags and possible cyclicalities. We shall, second, aim at understanding, or even perhaps at challenging, the measures in question. Analysis of the system defined by the model will then have to be made, with the idea that this system might well mimic, or perhaps deviate from, the mechanism or organism defined by the real economy.

This second approach will bring us back to question of forecasting accuracy raised in the foregoing section[178].

3.5.1. *Multipliers*

(i) The effects of changes in exogenous variables would be easy to trace if the model was linear, as was assumed in Sections 3.2 and 3.3. We would simply read these effects in the reduced form (177). For instance, a permanent change of size u occurring in period t on the exogenous variable x_i would have immediate effect v_{j0} on the endogenous variable y_j given by $P_{ji0}u$. In period $t+1$ the effect v_{j1} would be the j-th component of:

$$v_1 = [Q_1 p_{io} + p_{io} + p_{i1}]u, \tag{181}$$

where $p_{i\tau}$ would be the vector of the i-th column of P_τ. The long run effects $v_{j\infty}$ could be similarly computed by recurrence, leading in the limit to the j-th component of:

$$v_\infty = \left[I - \sum_{\tau=1}^{h} Q_\tau\right]^{-1} \left[\sum_{\tau=0}^{h} p_{i\tau}\right] u. \tag{182}$$

The vectors multiplying u in the right-hand members of respectively (181) and (182) would then characterize the intensities of the short-run (over two periods) and the long-run effects of a change in x_i. Formulas giving effects of a temporary change in x_i concerning just period t are similarly derived: for instance, deletion of the second term in the square bracket of (181) leads to the result for period $t+1$.

Actual macroeconometric models are not linear. This means that the intensities of the effects depend not only on the parameters of the system, like in (181) or (182), but also on the values of the variables $x_{i,t-\tau}$ and $y_{j,t-\tau}$ in the evolution with respect to which changes are assumed to occur. In other words the multipliers, defined above by the components of v_1/u and v_∞/u, can no longer be computed once and for all, irrespective of the path defining the base from which deviations have to be computed. But, once we know the model and a base evolution assumed to be ruled by the model, it is easy, using computations with the equations of the model, to simulate what the path would have

[178] About the analysis of macroeconometric models, see in particular M. Deleau et P. Malgrange, *L'analyse des modèles macroéconomiques quantitatifs* (Economica, Paris, 1978); R. Fair, *Specification, Estimation and Analysis of Macroeconometric Models* (Harvard University Press, 1984); K. Wallis, *Large-Scale Macroeconometric Modeling* (1995) (op. cit.).

been if the exogenous variable x_i would have taken, temporarily in period t or permanently from period t on, a value differing by a small amount u from its value or values in the base. Such *simulations* will directly give the relevant values of the multipliers if attention focuses on the base path and on its course from period t on.

When reporting values of multipliers we then have in principle to report the base with respect to which they were computed. In many cases it does not really matter because non-linearities have negligible impact on the results of the simulations examined, in comparison with the impact of sampling errors in the estimation of parameters. But it is not so for some multipliers, as we shall see in a moment. Therefore reporting results obtained from different reference bases is interesting, or simply from changes applied in different time periods presenting contrasting features. How to interpret different multiplier values found for the same couple of variables (i, j) may, however, raise questions if the references differ by many features, rather than by just one, and such is often the case.

(ii) The aggregate characteristic most commonly discussed is without question, "*the multiplier*", that is, the increase in real gross national product which results from a permanent increase of 1 in autonomous demand.

An examination of econometric models immediately leads us to make two distinctions. First of all, we should specify exactly what is meant by the "short-term multiplier", which usually designates the effect four quarters after the date of the increase in autonomous demand; in contrast the "long-term multiplier" concerns the limit effect after many quarters. In the spirit of a distinction made at the beginnings of Chapters 7 and 8, we often also find it appropriate to specify the degree of tension on productive capacities in the reference situation, since we expect the values of short and long-term multipliers to be sensitive to this degree of tension.

As an example, some results obtained in 1980 from the French quarterly model METRIC are set out in Table 2. There are noticeable differences between short and medium term effects; note also the incidence of the pressure on demand in the reference period; note, finally, the importance of the role played by foreign trade which was ignored in the theoretical models of Chapters 7 and 8. The increase induced in aggregate demand by an increase in government demand, is less important than was thought in some early Keynesian writings. Part of it is satisfied by increasing imports or by foregoing some exports, thus, without additional domestic production and with no generation of new income. Indeed, we notice in the table, that the impact on the balance of

Table 2
The effect of an increase in government demand according to METRIC[1]

	Fixed exchange rates				Flexible exchange rates and weak demand pressure[2]	
	Weak demand pressure[2]		Substantial demand pressure[3]			
Variation in	Short term[4]	Medium term[5]	Short term	Medium term	Short term	Medium term
Real gross national product (amount)[6]	1.5	1.7	1.5	2.8	1.7	2.0
Foreign trade balance (amount)[6]	−0.7	−0.4	−0.4	−0.9	−0.6	−0.2
Consumer price index (%)	−0.0	0.1	−0.1	0.5	−0.0	0.7
Industrial wage rate	0.0	0.5	0.0	2.1	0.1	1.2

[1]The amount of government spending on construction and public works was raised permanently by 1 billion francs per quarter at 1970 prices (that is 0.4% of gross domestic product in 1979 and 0.5% in 1971). The additional government deficit was financed directly for the Treasury by the Bank of France.

[2]The change occurs from the beginning of 1979, the reference period is 1979 to 1984.

[3]The change occurs from the beginning of 1971, the reference period is 1971 to 1976.

[4]After one year.

[5]After six years.

[6]Billions of francs at 1970 prices.

foreign trade is substantial in most cases[179].

This impact, and consequently also the multiplier, depend on the assumption made as to the reaction in the exchange rates. The latter are assumed to be fixed in the first part of Table 2. But it is more reasonable to suppose that, following the stimulation of domestic demand, the deterioration in the trade balance would provoke some fall in the exchange rates, which would improve

[179] The values shown in Table 2 for this impact assume that the stimulation of demand comes from government spending on construction and public works, which has a small import content. Higher values are found when the stimulation of demand comes from an exogenous increase in firms' investment in equipments. By comparison with the first column in Table 2, an increase of one billion francs produced, after one year and at fixed exchange rates, a deterioration of 1.0 in the amount of the trade balance rather than 0.7, and a GNP multiplier of 1.3 rather than 1.5.

international competitiveness, slow down imports and stimulate exports. Note on Table 2 that taking these reactions of the exchange rates into account, by adopting equations from the model METRIC, divides by two the medium-term impact on the trade balance and consequently raises the multiplier. The counterpart is, of course, an inflationary effect on the price level: permanent increase in public demand then produces, after six years, an increase of 2.0% in real gross domestic product and 0.7% in the consumer price level, this in reference to a situation where the pressure of demand would be weak.

As for the effect of pressure on productive capacities, Table 2 compares two time profiles concerning two different reference periods: a projection for the years 1979 to 1984, characterized by high unemployment, and the path observed from 1971 to 1976, a period in which a high rate of capacity utilization occurred during four years. (A lower tension on capacities during the later period is not the only difference between the two references; thus increased opening to foreign trade explains the larger short-term effect on the trade balance in 1979 than in 1971.) In the METRIC model increased pressure on capacity has mainly a direct effect on the growth rate of the wage rate, which has a stimulating effect on output but also an inflationary effect.

(iii) The overall impression we get from Table 2 is that of a smaller multiplier than that obtained from the elementary multiplier theory. We get a similar impression when we examine the many other macroeconometric models which were built. Already long ago B. Hickman[180] analysed 16 annual models, which were constructed before 1965 and concerned 10 different countries. Hickman called attention to the fact that, for 11 out of the 16 models, the short-term multiplier applying to the amount of government expenditure was between 1.23 and 2.06. The long-term multipliers tended to be higher but varied even more between the models. P. Artus and P.-A. Muet[181] examined in 1980 ten models built for the United States or Canada and five for France. They observed that the short and medium-term multipliers ranged from 1 to 2.3 for most of the models (assuming fixed exchange rates). The effects of increased public spending on prices were almost always weak in the short run; but they varied considerably in the medium run according to the models and according to the reference period considered.

A difficulty in such direct comparisons of "the multiplier" appearing in the analysis, by their authors, of various macroeconometric models comes from

[180] B. Hickman, Dynamic properties of macroeconomic models; an international comparison, in: M. Bronfenbrenner (ed.), *Is the Business Cycle Obsolete?* (Wiley, New York, 1969).

[181] P. Artus and P.-A. Muet, Une étude comparative des propriétés dynamiques de dix modèles américains et cinq modèles français, *Revue économique* (January 1980).

what they otherwise assumed in their simulations. This is apparent in Table 2 and was also well illustrated by R. Fair (1984, op. cit.). Presenting the results of simulations from his quarterly model of the US economy, Fair showed that a given permanent increase in government purchases of goods could lead to quite different results, depending on what was assumed about the money supply. Was the quantity of money unchanged with respect to its course in the reference path, as is usually assumed in IS-LM comparative statics analysis? Then the real GNP multiplier was equal to 1.2 in the short run (after a year), to 0.7 in the medium run (after four years). Was on the contrary the short-term interest rate kept unchanged, as is assumed in the elementary multiplier theory? Then the multiplier was 1.6 in the short run, 1.3 in the medium run, an increase in the quantity of money resulting from the stimulation of activity. Was it even the case that the extra government purchases were entirely financed by extra issues of central bank money, as was assumed in the French simulations reported in Table 2? Then the increase in the quantity of money was still higher and the multiplier was 3.2 in the short run, 2.3 in the medium run. The size of the differences shows the importance of monetary policy in macroeconomic dynamics. It also suggests the importance of strict standards in the comparative analysis of various macroeconometric models[182].

3.5.2. *Analysing the multipliers*

We might think that, by now, we ought to be able to draw more precise conclusions than we could in 1980. But the situation did not so much change in this respect from what it was then. Noting the fact might lead us to reflect on what seems to be a more general difficulty for the progress of empirical knowledge on economic phenomena; but it is not our subject at present. Let us rather comment on what is available and on how we might correct present deficiencies.

(i) A good documentation on values of the multipliers implied by the many macroeconometric models which were built ought, first, to be easily accessible to macroeconomists, hence more centralized than it now is. In order to be really instructive, this documentation should be organized by application of a

[182] After our earlier theoretical study of aggregate demand analysis in Chapter 7 we should not be surprised to find that, in macroeconometric applications, the numerical values of "the multiplier" may so much depend on the exact definition of this entity and on the nature of disequilibria in the reference situation. We saw in Section 2.3 of Chapter 7 the importance of what was assumed about the financing of a possible increase in government consumption, or equivalently about the impact on the monetary equilibrium. We saw in Part 4 of Chapter 7 how much the impact could depend on degrees of slack or tension on the markets for good and for labour. We saw in Part 5 of Chapter 8 how the picture was becoming still more complex when we were looking beyond the short term.

methodology which would aim at more than adhering to good librarian principles. It should also reveal the reasons why multipliers implied by different models differ. For the most part, what is available is not only limited in its coverage, but also little organized. Most often it seems to be made of simple juxtapositions of results presented by owners of models and obtained with little effort at standardization or interpretation: model-owners are left free to operate their models in the way they typically do; little attempt is made at analysing differences in the results.

K. Wallis[183] explains why such a practice does not suffice. He argues in favour of the alternative approach implemented since 1983 at the University of Warwick for models of the UK economy. Complete models and associated databases are deposited at the Macroeconomic Modelling Bureau in Warwick. The staff of the Bureau, a group independent of model-owners, is responsible for undertaking direct comparisons across models, according to the design, computations and analysis it finds best suited for explaining why simulations of the impacts following various shocks do not agree. The investigations of the Bureau particularly dealt with four quarterly models, all built in the income-expenditure tradition. They brought some interesting results.

The overall properties of the models differed, particularly so in the medium run for which the government expenditure multiplier varied from 0.4 to 2.2[184]. The Bureau suspected that this was related in particular to the fact that the models have different treatments of the exchange rate. The suspicion was supported by the fact that the important differences between the four models, in their assessment of the government-expenditure multiplier, practically disappeared when the real exchange rate was not allowed to deviate from its base-run values. But that was an irrealistic constraint, a fact which prompted the study of an exchange rate equation the specification of which would have theoretical merit and would cover the various specifications of the four models[185]. Careful estimation of the equation, and its substitution to the exchange rate equation in two of the models, reduced the differences in the model responses but did not eliminate them. Where it did not really help, attention had to be focussed on other areas of the model structure where alternative approaches had also been taken by the model builders.

One such area concerned the labour market. Again the Bureau suspected that different specifications explained, at least in part, why some multipliers

[183] K. Wallis, Comparing macroeconometric models: a review article, *Economica* (May 1993).

[184] P. Fisher, S. Tanna, D. Turner, K. Wallis and J. Whitley, Comparative properties of models of the UK economy, *National Institute Economic Review*, No. 129 (1989) 69–87.

[185] P. Fisher, S. Tanna, D. Turner, K. Wallis and J. Whitley, Econometric evaluation of the exchange rate in models of the UK economy, *Economic Journal* (December 1990).

were so different across models. Simulations had shown in particular different assessments of the impact on employment that would follow after five years from a given increase in government expenditure (assuming constant nominal interest rate): the percentage increase was estimated at 0.24, or 0.34, or 0.40, or even 0.79, according to the model used. It was then decided to look carefully at the labour market specifications of the four models[186]. It was found that the model giving the highest employment impact had unusual features in its representation of employment in non-manufacturing; a more acceptable econometric specification lowered the estimate of the impact to 0.53. The models also differed in the relationship between changes in unemployment, employment and the working population; the Bureau found that one specification was preferable on economic and statistical grounds; introducing it in the two models that did not use it also contributed to reduce discrepancies. Finally two of the models had incorporated the feature that cuts in income tax would lead to lower nominal earnings growth; this had considerable bearing on the properties of these models; finding that statistical evidence for inclusion of the feature in question was weak, the Bureau again revised the specification of the two deviant models; this considerably reduced the medium-run multiplier of a cut in the income tax, thus bringing the results more into line with those given by the other two models.

All things considered, these investigations show that important differences in model properties can often be resolved.

(ii) At this point it may be interesting to report some French work, done twenty years ago, led less systematically but inspired by the same concern about understanding how the specification of various parts of a macroeconometric model may affect its overall properties.

Two models were used in France in the 1970s for macroeconomic policy analysis: METRIC and DMS. They were giving in some cases significantly different results. It was then important to have understood that this usually came from the fact that firms' investment was modelled differently. The DMS model attributed a substantial role to realized and anticipated profit margins, whereas the METRIC model was in contrast exhibiting the influence of the relative cost of capital with respect to labour (both models were, of course, also recognizing the part played by the accelerator)[187]. Motivated by the realization

[186] D. Turner, K. Wallis and J. Whitley, Differences in the properties of large-scale macroeconometric models: the role of labour market specifications, *Journal of Applied Econometrics* (October–December 1989).

[187] For an example of how such differences were taken into account in policy analysis, see G. Oudiz, E. Raoul and H. Sterdyniak, Réduire la durée du travail, quelles conséquences?, *Economie et Statistiques* (May 1979).

of the problem, the authors of the DMM mini-model aimed at finding out exactly what the most frequently used multipliers depended on. It is interesting to examine some of their conclusions now.

First of all, the authors (Deleau, Malgrange and Muet, op. cit.) turned attention to how the real block of the hierarchical model functioned, that is, the model defined by Equations (150) to (158). This real block was likely to provide a good approximation of short-run effects, since effects on prices and wages were recognized to be rather slow to appear. Let us look at the classic multiplier, defined as the ratio $\delta y/\delta G$ between the change in the volume of output and the change in government demand, other exogenous variables remaining constant. We can see immediately that, in the short run, this multiplier is a decreasing function of λ_2, λ_3 and λ_5, hence an increasing function of the speed of adaptation in real demand variables. Thinking in terms of annual periods we can consider the following values of the parameters as representative: $\lambda_2 = 0.5$; $\lambda_3 = 0.7$ and $\lambda_5 = 0.5$; using plausible values for the other parameters we thus arrive at a short-term multiplier of the order of 1.6.

The multiplier is sensitive to changes in λ_2 and λ_3 mainly through their effects on investment and to changes in λ_5 through their effects on consumption; for the multiplier to be greater than 2, for example, it suffices to replace the above values by $\lambda_3 = 0.5$ or by $\lambda_5 = 0.3$. Since consumption function estimates are less uncertain than investment functions, uncertainty about the value of the short-term multiplier is likely to come from the investment equation. Indeed, the effects after a year, as estimated from the two French models METRIC and DMS with respect to the same reference path, showed the following values for the multiplier: METRIC 1.38 and DMS 1.1. But this corresponded to the same value of 0.31 for $\delta C/\delta G$ whereas $\delta I/\delta G$ amounted to 0.66 for the first model and 0.31 for the second.

We can now study the effect of a demand stimulation δG on the price level by using the price-wage block of the hierarchical model corresponding to *DMM*, that is Equations (159) to (161). After one period, two parameters play an important role: λ_7 characterizing the inertia of prices with respect to costs and the coefficient f_2 capturing the role of the degree of capacity utilization. Depending upon whether f_2 is low or high, the first term in the bracket of Equation (161) will dominate the second, or not; this first term decreases when δG is positive because of the favourable effect on the productivity of labour. On the other hand, after three or five periods (three or five years), what will happen also depends very much on two other coefficients. The more complete the indexation of wages to prices (g_2 is higher) and the stronger the Phillips' effect ($-g_3$ is higher), the more noticeable the inflationary consequences. Since the appropriate values for the four coefficients λ_7, f_2, g_2 and g_3 must depend on the country, the period and the level of economic activity in the reference path, as we saw in Chapter 8, we can explain the very different medium-run inflationary effects predicted by various models as resulting from the same policy of demand stimulation.

But the main advantage of the DMM model might be that it allows to study different feedbacks from prices, costs and the monetary equilibrium to the volume of production, that is, to study the different channels of interaction between monetary and real phenomena. Thus, we can consider:

- the effects of international competition: since θ_2 is positive and θ_4 is negative, the stimulating effect of δG will be reduced in the medium run by an increase in imports and a fall in exports;
- the income distribution effects, through changes in the labour share;
- the wealth effects, an increase in prices reducing the purchasing power of assets and real burden of debts (effects which were ignored in the DMM model);
- the effects through those relative costs and relative prices which might affect firms' decisions;
- the liquidity effects: tension on the money market can reduce the stimulating effect of δG as a consequence of increased interest rates and cost of capital, or as a result of credit rationing.

The discussion clearly shows that in the medium run all these effects may play a part, especially when the demand stimulus has a significant impact on inflation. It is not possible to take up here the detailed study of M. Deleau, P. Malgrange and P.-A. Muet. We just note their conclusion, namely that many different long-term developments can result from the interaction between monetary and real phenomena, but that they are likely to have a moderate force and to play relatively little during the first few years following an exogenous change.

3.5.3. *Cyclical properties*

Macroeconometric structural models should answer questions about the possible cyclical features of the evolution of economies to which they apply. Such questions were indeed raised in the context of theoretical models since the beginning of this chapter. We then asked: might the solutions to the system defined by the set of behavioural equations and equilibrium constraints exhibit cycles? If so, are these cycles explosive or damped?

Our examination thus far brought out clearly the multiplicity of factors which may come into play and the uncertainty which remains if we adopt a purely theoretical approach considering the various factors consecutively. This was apparent already in Part 1 of the chapter, when we tried to build around a reference cycle, which was supposed to have its origin in the accelerator: even the very existence of this underlying cycle could be disputed since it was proved only for particular specifications of the very simple model then considered. More generally, the number of relevant factors, the different ways in which each of them can appear, the sensitivity of the resulting time paths to the force of each factor and the lags before it takes full effect, all this explains why no definite conclusion can be drawn without the help of econometric studies which treat all the main economic variables characterizing business cycles simultaneously. It was thus very natural to examine the cyclical properties of existing macroeconometric structural models.

For a time hopes were entertained that such properties would exhibit enough similarity across models for being instructive about the cyclical behaviour of our economies in the absence of exceptional shocks. But it turned out that results did not meet such hopes: the medium and long-run time paths generated by simulations from various models differed too much. No business cycle theory was emerging from the econometric work on structural models.

The first investigation in this respect was due to I. and F. Adelman[188] studying the Klein–Goldberger model (op. cit.). They found a dominant cycle with a period of three to four years, but with a damping factor quicking reducing the amplitude of the fluctuations (this strongly damped cycle was due to the income distribution effects exhibited in Section 1.5 of this chapter). More systematic studies showed that even this limited result did not generally apply: the quarterly American models existing around 1970 did not display interesting cyclical properties, unless for some models convenient serially correlated disturbances were introduced[189]. The diagnosis was sufficiently negative for cyclical properties to be apparently ignored in more recent writings about structural macroeconometric models, such as R. Fair (1984) or K. Wallis (1995) op. cit.

The question to know what could be the explanation of these disappointing results was investigated by the authors of the DMM model (Deleau et al., 1984, op. cit.). They looked at whether the parameters of their mini-model could be chosen in such a way that the paths generated by the model would exhibit cyclical features. They concluded, first, that such features had to come essentially from the real block of the model: although the various channels of interaction with other blocks allowed additional cyclical components to appear, those

[188] See I. Adelman and F. Adelman, The dynamic properties of the Klein–Goldberger model, *Econometrica* (October 1959); see also E. Howrey, Stochastic properties of the Klein–Golberger model, *Econometrica* (January 1971).

[189] B. Hickman (ed.), *Econometric Models of Cyclical Behavior*, N.B.E.R. Studies in Income and Wealth (Columbia University Press, 1972).

seemed to have periodicities which were too long to be relevant (twenty years for example). The real block of the hierarchical model could lead to a cyclical path only if investment (which should be understood as concerning in particular inventories) reacted quickly to changes is expected output y^*, (the result is consistent with the study of the accelerator made in Section 1.2). We see, for example, in their results that a cyclical component of some considerable amplitude appears when a value of 0.3 is given to coefficient λ_3 of Equation (155) (a value actually much lower than that preferred by the authors of the model: $\lambda_3 = 0.7$). The period of the damped cycle is then about eight years. The amplitude is reduced to a fifth from one cycle to the next.

What are the implications of this experience for business cycle theory? We shall not try here to really answer the question, but rather warn against the interpretation which would say that actual business cycles would quickly disappear if the environment remained unchanged for several years. For such a conclusion to be founded, the existing models would have to provide a faithful representation of the real economic dynamics, whereas they are exposed to substantial specification and estimation errors.

The specifications chosen appear rather crude when one considers the many factors which can influence short-run evolutions. Each behavioural equation is simpler than those obtained from specific studies concentrating on the same behaviour. For instance, cyclical features may have disappeared because of an incorrect representation of the accelerator, whose important role for inventories and for industrial investment is well known. Or lags in private or public decisions may have been misspecified.

Reference to lags calls attention to estimation errors, which may be important because the data base is made only of aggregate time series on a limited period (econometric methodology shows clearly that a precise estimation of relationships including distributed lags would call for a quite large sample). More generally the data base does not give enough information to allow precise estimation of the various behavioural equations.

The inevitably approximate character of existing models is a particular nuisance when we are trying to determine whether spontaneous unperturbed evolutions would be stable or unstable. In fact, stability is as sensitive to the specification of the model as it is to the estimation of coefficients and the lags included in the equations. This is indeed the reason why different models often predict time paths which considerably diverge from one another after the first few years.

All in all, we have to realize once again that we still do not know very much about the dynamic characteristics of our economies. None of the simple schemes presented in this chapter should be taken as established truth. Each of them serves more as an aid to understanding the nature of business cycles than as a full description of the mechanisms that generate them.

However, it would be a mistake to think that accurate knowledge of cyclical features is a necessary prerequisite for wise countercyclical policies. Macroeconomic policies aim at avoiding some unwanted effects of exogenous shocks or at correcting tendencies leading in directions which are obviously unfavourable. Their objective is short-term regulation rather than complete and definitive stabilization of the economic system. They certainly have to take into account the recognized non-immediate effects. But they will act in the future again. The proper image is that of a driver piloting a car, without knowing exactly the profile of the road ahead. In the hands of a well educated economist, the macroeconometric model, with all its limitations, ought to be a useful instrument during such a piloting.

3.5.4. *Evaluation of models*

Practically since the beginning of this section we repeatedly touched on questions of evaluation, as we were analysing the properties of macroeconometric models and trying to be in a better position to understand the origin of these properties. But we should like to now take up evaluation more directly and more globally. How good are models at predicting the evolution and reactions of the economy? The evolution, when the purpose is forecasting, the reactions to exogenous changes, when it is analysis of alternative policies. The question is phrased in relative terms, as was in Section 3.4 the question posed about the accuracy of macroeconomic forecasts as usually practiced. Indeed, part of the issue is to compare predictions that a model would have given to observation of what is actually happening. Comparing the predictive performances of different models is also part of issue.

In order to deal with our present subject we shall mainly consider its discussion in the book published in 1984 by Ray Fair (op. cit., see in particular its Chapter 8)[190].

(i) In the 1970s the standard methods for model evaluation were focusing on *ex post* forecasts. The predictions, which would have been given by the model for the main endogenous variables if the forecaster had known the subsequent values of the exogenous variables, are iteratively computed outside the sample period that served for estimation. Comparison with the actual series of those same endogenous variables gives the forecast errors, from which a measure of forecasting accuracy can be derived. The *ex post* forecast errors computed

[190] Readers may also be interested in Y. Chong and D. Hendry, Econometric evaluation of linear macroeconomic models, *Review of Economic Studies* (August 1986).

over a common simulation period were also used for comparison of different models. Besides a number of details in the application of this approach, two points have a sufficiently general interest for us to consider them here.

In the first place, and this applies to policy analysis as well as to forecasting, use of an econometric model most often rests on a *deterministic* simulation, in which a single point value of each of the variables of interest is calculated. The endogenous variables of an econometric model are random variables, however, and for various purposes it is useful to have more complete knowledge of their probability distributions. The common method, the only one workable with non-linear models, consists in using a *stochastic simulation*, in which many draws of the disturbance terms appearing in the equations are made and as many solutions of the model are calculated, the statistical distribution of the values found for a variable being then an estimate of its probability distribution. For model evaluation stochastic simulations are obviously preferable to deterministic simulations.

In the second place, for model comparison, the presence of exogenous variables creates a serious problem. Models differ both in the number and types of variables that are taken to be exogenous and in the sensitivity of the predicted values of the endogenous variables to the exogenous-variable values. The common procedure of model comparison does not take these differences into account. If one model is less "endogenous" than another (say that prices are taken to be exogenous in the first model but not in the second), it has an unfair advantage in the comparison of errors as commonly measured. This does not much matter in comparisons between models having similar endogeneity, such as the four UK models referred to earlier here. But it does in comparison between forecasts made respectively by a structural model and by a "naïve" method.

Fair explains in his book how he deals systematically with the four main sources of uncertainty of a forecast: (1) the disturbance terms, for which stochastic simulation is the answer, (2) the estimates of the coefficients of the model, for which stochastic simulation must be applied to the sample period and the resulting estimations, as well as to the forecast period, (3) the exogenous-variable forecasts, which we shall now consider, (4) the possible misspecification of the model, about which we shall say a little more later.

Two polar assumptions can be made about uncertainty of the exogenous variables: there is no uncertainty or, at the other extreme, forecasts are in some way as uncertain for exogenous variables as for endogenous variables. While the first assumption, which underlies *ex post* forecasts, is clearly underestimating exogenous-variable uncertainty, the second assumption is likely to overestimate it. This is particularly true for fiscal policy variables in structural models, where government budget data are usually quite useful for forecasting over a number of quarters ahead. In his comparisons Fair carefully avoids giving his own model an unfair advantage: in order to estimate the degree of uncertainty attached to forecasting an exogenous variable he relies on a fit of this variable to an eight-order autoregressive equation with a linear time trend. It turns out that, even though probably overvalued, this source of uncertainty in his model does not play a large part: for real GNP it is negligible in forecasts made two or three quarters ahead, it amounts to 0.25 in the estimated 3.43 per cent standard error of forecasts made eight quarters ahead.

(ii) Fair evaluated the predictive accuracy of his quarterly model of the US economy and of four alternative quarterly econometric instruments for the same economy. We begin with the root mean squared errors of outside-sample forecasts (which means forecasts obtained in quarter t from the instrument fitted to past quarters up to $t - 2$). The measure is computed over forecasts made for quarters running from 1970I to 1982III. It is given for five macroeconomic variables and for eight forecast horizons (1 to 8 quarters ahead). Some macroeconomists may think that longer forecasting horizon would have been interesting also.

Reading his Table 8–3, we find that in all cases predictive accuracy quickly decreases with the horizon (forecast uncertainty increases): for instance, his model's prediction of real GNP has a RMSE of 0.8 per cent one quarter ahead, 3.8 per cent eight quarters ahead; the difference is least marked for the short-term interest rate (from 1.1 percentage point to 2.6). For his model the worst performance may appear to be in the prediction of the quantity of money (RMSE increasing from 1.3 to 6.6 per cent).

Fair's model of the US economy is fairly large (30 behavioural equations, 98 accounting identities). A first benchmark for comparison is a mini-structural linear model based on a very crude theory, containing eight behavioural equations and four accounting identities (LINUS). For only two of the five macroeconomic variables can the comparison with the large model be made. For the short-term interest rate the mini-model is almost as good as the full model (RMSE from 1.2 to 2.9 percentage point when the horizon shifts from one to eight quarters ahead). For real GNP the performance of the mini-model quickly decreases with extension of the forecast horizon (RMSE from 0.9 to 6.0 per cent).

The "naïve forecast" alternative is obtained by independent projections of each variable thanks to an autoregressive equation, with eight lags and a linear time trend, fitted on the observed series during the sample period. This naïve instrument has an overall surprisingly good performance, in particular when compared with the two VAR instruments, to which it looks to be superior in all respects except for forecasting the GNP deflator more than two quarters ahead. Fair's full model does better for three of the five variables; for instance, the RMSE of the naïve forecast of real GNP increases from 1.2 to 4.3 per cent with the forecast horizon. For the GNP deflator it is still mildly superior at horizons exceeding four quarters, but else practically equivalent. Finally, for the quantity of money the naïve forecast (RMSE from 1.2 to 3.4 per cent) definitely beats that given by the large model.

The two vector-autoregressive models concern the five macroeconomic variables in question: VAR 1 explains the vector of these variables by its first six lagged values and by linear time trends; VAR2 saves on the number of coefficients to be estimated by reducing to two for each variable the number of lags with which the other four variables appear in the equation with this variable in the left-hand member. VAR 2 is doing better than VAR 1 for four of the variables, but the results are not so different. The VAR models are beaten by Fair's model except for the quantity of money and for the GNP deflator.

(iii) Ray Fair wanted to analyse such results in order to see what part each of the four main sources of forecast uncertainty could have taken in forecast errors. Stochastic simulations gave the parts due respectively to the disturbance terms in the equations and to errors in estimation of the coefficients. The part coming from errors in the forecasts of exogenous variables was analysed as indicated earlier in this section. As for the part coming from possible misspecification of the model, Fair explains how he evaluated it by the difference between two statistics estimating respectively: the full variance of the forecast error, and the variance resulting from the preceding other sources of uncertainty combined. Computations for this analysis concerned forecasts for 1978-I to 1979-IV.

The results show that two parts make only small contributions to the overall uncertainty of US forecasts made by the five models examined: the parts coming from errors in respectively the estimation of coefficients and the forecast of exogenous variables. The other two parts shared the main responsibility, misspecification appearing as the most important. It will be enough to consider here the forecasts, eight quarters ahead, of real GNP, the GNP deflator and the quantity of money, as they result from Fair's macroeconometric model.

For real GNP the standard error of the forecast would amount to 1.3 per cent if only disturbance terms mattered. It would increase to 1.6 if, the effects of uncertainty in coefficient estimation and exogenous variable forecast being taken into account, no misspecification would have occurred. All four sources of uncertainty being taken into account, the standard error of the forecast really amounts 3.4. The corresponding three figures are (0.9, 1.1, 4.3) for the GNP deflator and (2.0, 2.3, 5.3) for the quantity of money. The picture is similar when errors following from the other four forecasting instruments are analysed. In the conclusion of his book, Fair writes: “All the models tested appear to be misspecified by a fairly large amount” (p. 409).

(iv) Considering only RMSE of forecasts in comparisons of macroeconometric models may not provide adequate answer to some relevant questions such as: how should we interpret the differences in forecasts? Does each model have a strength of its own, so that each forecast represents useful information unique to it, or does one model dominate in the sense of incorporating all the information in the other models and more?

Such questions are particularly relevant in the comparison between large structural econometric models and the much smaller VAR models. Structural models with their numerous exogenous variables use large information sets, but rely on identifying restrictions which cannot claim to be exactly true. Can we say, at the end of the forecasting process, that they have carried useful information ignored in the VAR models?

R. Fair and R. Shiller devised a test for answers to the above questions and applied it to model-based forecasts of the US real GNP growth rate[191]. Some explanation of the test is necessary before we look at the application. Let y_t be the observation in period t of a variable of interest, let ${}_{t-h}\hat{y}_{kt}$ be the forecast that in period $t-h$ model k delivers for the value y_t, by application of the procedures specific to owners of model k. The Fair–Shiller test is meant to assess the additional information that the model $k=2$ could bring when y_t has been forecast by model $k=1$, additional information which would be useful for improving upon the forecast in question, and *vice versa* the additional information that model 1 could bring when y_t has been forecast by model 2. The test is defined in the framework of an assumed regression equation:

$$\begin{aligned} y_t - y_{t-h} &= \alpha + \beta\left({}_{t-h}\hat{y}_{1t} - y_{t-h}\right) \\ &\quad + \gamma\left({}_{t-h}\hat{y}_{2t} - y_{t-h}\right) + \eta_t, \end{aligned} \tag{183}$$

where α, β and γ are parameters and η_t is an error, the mathematical expectation of which is equal to zero. The proposed procedure is, after estimation of the parameters α, β and γ, to test $\gamma = 0$, and *vice versa* $\beta = 0$.

This procedure raises a number of econometric problems, which will not be considered here. They concern the specification of Equation (183) and the exact definition of the test. Let us simply record that reference to equations similar to (183) has been a common practice in the forecasting methodology

[191] R. Fair and R. Shiller, Comparing information in forecasts from econometric models, *American Economic Review* (June 1990).

(see K. Wallis (1995, op. cit.) for this). As for the definition of the test, it has to take account of the fact that heteroskedasticity and serial correlation of the error process $\{\eta_t\}$ is to be most often expected. But a test, taking that into account and using the ordinary least squares estimates of the parameters, is defined by Fair and Schiller.

The test is applied to one-quarter-ahead and four-quarters-ahead forecasts obtained from a 1976-version of Fair's structural model and from three kinds of models without exogenous variable. In the same way as in the work of Fair reported above and for the same reason, the forecasts based on the structural model use projected values of each exogenous variable which extrapolate an eight-order autoregressive equation (with a linear time trend) fitted on the past series of this variable.

As alternative models we find: first, two naïve autoregressions of real GNP, with respectively four and eight lags, and a linear time trend; second, six VAR models, all involving three other variables treated simultaneously with GNP, the differences between these VAR models concern either the number of lags in the regression, or whether and how Litterman's "Bayesian" device is applied; third, eight "autoregressive components" models (*AC*) meant to show whether the paths followed by disaggregated demand components of GNP contained useful information. Forecasts from a AC model are obtained by simple aggregation of independent forecasts of the components, each based on an eight-order autoregression with a linear time trend. The AC models differ by the degree of disaggregation (from 6 to 48 components) and by whether or not the first four lagged values of real GNP are added to each autoregression.

Results show that the structural model leads to the best forecast in terms of RMSE. They also strongly support the hypothesis that this model carries useful information not carried in the other models (γ significantly differs from zero, but not from one, when the structural model plays the role of model 2 in Equation (182) with $y_t = \log(\text{real GNP})$).

However, a t-statistics of 1.9 is close to rejection of the hypothesis that $\beta = 0$ with any of the VAR models used as model 1 for four-quarters-ahead projection. Hence, VAR models contain useful information not captured by the structural model. This suggests that the dynamic specification of the latter model could be improved.

It is also interesting to note that information about demand components, as captured in the *AC* models, does help improve forecasts. The conclusion emerges when the *AC* models are compared with the VAR models. It shows that, at a high level of aggregation, a VAR model misses detailed information that is useful even for prediction of the most aggregated variables.

Ending here this subsection on the evaluation of models, where Fair's structural model fared relatively well, we should repeat the warning given by Fair and Shiller in their conclusion: "The fact that one model does well or poorly for one sample period (in our case 1976III–1986II) does not necessarily mean that it will do well or poorly in future sample periods".

3.6. The framework of policy analysis

From its inception aggregate demand analysis was meant to apply as much to policy analysis as to forecasting. Macroeconometric models were built in order to serve both purposes more realistically than simple theoretical models could do. A methodology for policy analysis then emerged, based on a quite natural logic. It was progressively enriched as experience in the use of structural models was requiring more precise guidelines, as the ideas embodied in the models were becoming more complex, as reflections on the methodology itself was pointing to some weaknesses of the practice which was commonly applied and recommended. In this evolution a landmark was the "Lucas'critique"[192] arguing that policy analyses made with structural models were subject to a basic flaw because they neglected the impact of public decisions on private expectations about future policies. How to cope with the critique will be a major theme in the next part of the chapter. Keeping to the announced organization of the chapter, we shall in this section and the next accept the neglect in question in our study of the methodology. This will permit a less heavy entry into our present subject than would have otherwise been the case.

The next section will deal with the dynamic formalization of the methodology. We shall then even accept the context of what now appears to have been a climax of Keynesian ambitions: the optimal control of economic evolution. This section will prepare the ground in three ways: by a heuristic discussion of what can be expected from a macroeconomic strategy for the stabilization of business cycles, by the definition of a formal framework, by the characterization of optimal static policies.

3.6.1. *The challenge confronting stabilization strategies*

The ideal for a macroeconomic strategy would be to ensure permanent full employment with no inflation and with no rationing on any market. But to

[192] R. Lucas, Econometric policy evaluation: a critique, in: K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, Supplement to the *Journal of Monetary Economics* (January 1976).

achieve completely this ideal objective turns out to be impossible for three reasons whose effects reinforce each other.

In the first place, any counter-cyclical policy takes longer to have its full impact than was often thought. There is always a period of a few months between the time an action is decided and the time the effects are felt, usually several quarters. According to the normal procedure national public budgets are prepared some six months before the beginning of the year they concern. Fiscal policy can take exceptional measures within the framework of a "plan to boost the economy" or a "plan to curb inflation". But even if authorities speed up the application of their decisions, delays are unavoidable in the allocation of new credits or budget cuts, in the repercussions on orders to suppliers or on recruitment of staff. Delays are also unavoidable in the distribution of new subsidies or the collection of new taxes, or still in implementing the decision to accelerate or stagger the realization of public investment projects.

Monetary policy seems to be more flexible. But there are also lags in the effects, already most often between the time the central bank reorients its policy and the time that firms and households see a real change in the cost or availability of credits to which they have access. And this concerns just the first impact of the measures taken. There are other lags, usually longer, before new orders are executed, new incomes spent or private plans revised. The multiplier effects only start afterwards.

In the second place, both behaviour and the economic environment are subject to erratic and unforeseeable changes. Hence, economic policy cannot be perfectly tailored in advance so that the consequences of policy measures would always make themselves felt at the right time and in the right place. Certainly, policy makers know that the effects of their decisions will spread over a number of quarters; they so consider available macroeconomic forecasts. But these cannot be completely accurate, particularly because the diagnosis will often have to rely on very approximate models. The measures chosen will hardly ever prove to be completely adequate since the situation observed will almost always differ from the one that had been expected.

In the third place, to these sources of inaccuracy, which are in the nature of things, we must add the effects of imperfect statistical knowledge. Statistical information is never complete nor instantaneous. As a consequence of gaps in information, observers may not be aware of some inventory building which will later require destocking and hence, produce a sharp fall in aggregate demand. As a consequence of lags in information, observers may be slow to diagnose the turning point of a cycle which would demand rapid corrective measures. Lags in diagnosis are added to lags in decision making and to delays in the effect produced by these decisions.

These conditions, in which the conduct of economic policy takes place, present a danger for governments who are in a hurry to see the results of their policy.

We had occasions to observe that, after the late diagnosis of a business downturn, some initial measures were adopted which went in the right direction but which, for several months, produced no perceptible effect; a second, then a third train of measures going in the same direction were adopted, and the latter usually stronger and stronger. The upturn in activity occurred in part as a result of the first measures, but it developed its own dynamics, which was at first favourable, and then progressively less and less satisfactory. The last expansionary measures to have been put into operation had their main effects in totally reversed conditions and amplified the business cycle which they were meant to dampen. Such clumsiness in the conduct of economic policy, which was called *stop and go*, characterized the way some governments acted in the late 1960s and the 1970s.

This bad experience still provides arguments for those who preach up, and have often done so for a long time, a much less interventionist macroeconomic policy. According to them, the intrinsic path of the economy is naturally stable, and short-term policy acts too blindly to be beneficial. Governments ought to establish a long-term line of conduct and have the strength not to deviate from it regardless of current information. The line of conduct is usually conceived as having to include strict equilibrium of public budgets and a moderate growth of the money supply, at a constant and predetermined rate. According to promoters of this thesis, inflation of any sort would exhaust itself in the absence of liquidity if the supply of money continued to grow at the predetermined rate; likewise, an increase in unemployment should be treated as accidental since this, too, would be reabsorbed spontaneously.

There is no doubt that the above thesis is too extreme. Simply ignoring current information about business conditions, and the changes it reveals in behaviours or in the economic environment, cannot be the optimal strategy. We discussed the issue in Chapter 8 and shall come back to it. The problem, however, remains to know just how sensitive economic policy has to be to accidental changes in current trends.

Those who consider that economic theory is uncertain, econometric models rather imprecise, statistical information questionable or out of date, want to see greater prudence with regard to deviations from the financial rigour which is justified in the long run. Even those with a less negative attitude towards the value of knowledge as a basic for action, can find in the way in which the economic system functions, reasons for recommending restraint in the handling of the macroeconomic policy instruments.

Indeed, our ultimate goal is the study of the following question: taking into account the laws governing business fluctuations and the means which public authorities have at their disposal, what is the best stabilization strategy? This section and the next one concern methodological preliminaries, which have a more general relevance and could have come earlier in this book.

3.6.2. *Models, instruments, informations*

The basic methodology of economic policy is fairly straightforward and was well exposed by Jan Tinbergen in the 1950s[193]. We shall take up its various elements as they can be presented now.

(i) From the first chapter onwards, we have distinguished, in any economic model, between the endogenous variables supposed to be determined within the model and exogenous variables which have to be present in the explanation but are assumed to be determined elsewhere. Among the exogenous variables we now have to systematically identify the "instruments" of economic policy, equivalently called the "decision variables". The other exogenous variables are "data" on which this policy does not act but which it must take into account. There is some hesitation in the literature about the designation of these variables. Sometimes they are just called "exogenous", it being understood that decision variables are then excluded. They have sometimes been called "state variables", but the latter name is normally meant to have a larger coverage: the state variables in period t are not only the exogenous variables in question, but also the past values of all variables, including the endogenous and decision variables. We may speak here of variables "on surrounding conditions". The expression "forcing variables" is also used, referring to the idea that the values of these non-decision exogenous variables put constraints on the choice of the values of decision variables, but again the expression is also used in a broader sense covering, for instance, all exogenous variables and random disturbances. In any case these other exogenous variables are meant to be observed in the current period.

In addition, we must recognize that economic phenomena are almost always influenced by causes other than those represented by exogenous variables. Sometimes theses causes are not even known, sometimes they are not observable, sometimes still there are too many additional causes and, taken individually, they have too little influence for featuring in the model. To account for the action of these unknown or unidentified causes we introduce into the corresponding equations terms which are not meant to be observed. These

[193] J. Tinbergen, *On the Theory of Economic Policy* (North-Holland, Amsterdam, 1952); *Economic Policy: Principles and Design* (North-Holland, Amsterdam, 1956).

terms are often called “disturbances”, but here we shall speak of “errors”, the expression most commonly used in the econometric literature.

Let x_i $(i = 1, 2, \ldots, m)$ be one of the m exogenous variables on surrounding conditions, its value taken during period t being denoted by x_{it}. Likewise, y_{jt} will refer to the value of the endogenous variable y_j $(j = 1, 2, \ldots, n)$ and z_{kt} to that of the instrument variable z_k $(k = 1, 2, \ldots, p)$.

The model used in the study of the economic policy in question determines the sequences of the values of the endogenous variables y_{jt} from the sequences of the exogenous x_{it}, the instruments z_{kt} and some errors. For simplicity we shall stick in our formalization to the case of purely linear models, including as many equations as endogenous variables. The j-th equation will be written as:

$$\sum_{l=1}^{n}\sum_{\tau=0}^{h} b_{jl\tau} y_{l,t-\tau} = \sum_{i=1}^{m}\sum_{\tau=0}^{h} a_{ji\tau} x_{i,t-\tau} + \sum_{k=1}^{p}\sum_{\tau=0}^{h} c_{jk\tau} z_{k,t-\tau} + \zeta_{jt}. \tag{184}$$

This equation exactly corresponds to Equation (174) except for minor changes in the notation and for the fact that the two kinds of exogenous variables are now distinguished. We shall not repeat here what was explained in Sections 3.2 and 3.3. Let us simply note that lags of any length up to h can be taken into account. More importantly we remember that many of the structural coefficients $b_{jl\tau}$, $a_{ji\tau}$ or $c_{jk\tau}$ are zero, or have a priori known values, and that this permits identification. In Equation (184) the error term ζ_{jt} is zero for accounting identities. If not, it is meant to sum up influences of non-observed or unidentified causes. The system of the n equations similar to (184) is more conveniently written in matrix form as:

$$\sum_{\tau=0}^{h} B_\tau y_{t-\tau} = \sum_{\tau=0}^{h} A_\tau x_{t-\tau} + \sum_{\tau=0}^{h} C_\tau z_{t-\tau} + \zeta_t. \tag{185}$$

Since the system is supposed to show how the y_{jt} are determined, it can be solved with respect to these variables, that is to say the matrix B_0 is non singular, and we can write:

$$y_t = \sum_{\tau=0}^{h} B_0^{-1} A_\tau x_{t-\tau} + \sum_{\tau=1}^{h} \left(-B_0^{-1}\right) B_\tau y_{t-\tau} + \sum_{\varsigma}^{h} B_0^{-1} C_\tau z_{t-\tau} + \varepsilon_t, \tag{186}$$

where ε_t is the new vector of errors, equal $B_0^{-1}\zeta_t$. The system of n Equations (186), also called the "reduced form", determines the value y_{jt} of each endogenous variable as a function of an error, of current or lagged values of the exogenous variables and of lagged values of the endogenous variables; all these values are also said "predetermined".

(ii) To make use of the system, in the form (185) or (186), we still have to *specify the information the decider has* at the time he chooses the values z_{kt}. We shall make in this respect various assumptions depending on the problem discussed, but they will always be very simple; we have to appreciate just how simplified they really are.

We shall often proceed as if the values of all coefficients were known exactly. This is never the case in reality since many of the coefficients are obtained from econometric estimates, the results of which are never perfectly precise. To get away with this difficulty we shall incorporate the effects of estimation errors into the effects of unidentified unknown causes. Hence we shall argue as if all these effect were involved in the values of the ζ_{jt} or ε_{jt}.

We shall also often assume that these errors are random variables and that we know the probability distribution of the sequences of these random variables (that is of the stochastic process of the ζ_t or the ε_t). Or rather we shall then assume that the relevant characteristics of this distribution are known. To justify the assumption it can be said that econometric procedures allow us to estimate these characteristics. Alternatively we can adopt a "Bayesian" point of view and say that the decider will have to choose; for this purpose he or she will have to know which probability distribution to attribute, in one way or another, to the errors. It is this "subjective probability" which is meant to enter into our model.

Finally, we are going here not only to make the two preceding assumptions but also to assume that the decider knows all the values of the $x_{i,t-\tau}$ (for $i = 1, 2, \dots, m$; $\tau = 0, 1, \dots, h$), as well as all lagged values of the endogenous variables, that is the $y_{j,t-\tau}$ (for $j = 1, 2, \dots, n$; $\tau = 1, 2, \dots, h$). This is an important simplification in both cases. Observing data always demands a certain amount of time, so the decider has to estimate some $x_{i,t-\tau}$ from incomplete information and often he or she even has to frankly forecast some other $x_{i,t-\tau}$, especially the x_{it}. Observation of endogenous variables also includes lags, which differ for different variables: some $y_{j,t-1}$ can still be unknown at the time when the z_{kt} are chosen.

It is possible to transform system (184) so that the last assumption is approximately satisfied. We can replace the still unknown vector $x_{t-\tau}$ by the sum $x_{t-\tau}^e + \xi_{t-\tau}$, in which $x_{t-\tau}^e$ is made up of whichever estimates or forecasts of the $x_{i,t-\tau}$ are available at time t when decisions about the z_{kt} are taken; the terms arising from the estimation or forecast errors $\xi_{i,t-\tau}$ can be merged with the ε_{jt}. Likewise, if the observation of a value $y_{j,t-1}$ or $y_{j,t-2}$ is not available, it can be replaced in the right hand side of (186) by its expression given by system (186) itself, written to calculate the vector y_{t-1} or y_{t-2}. (If, for example, (185) leads to $y_t = by_{t-1} + \varepsilon_t$ with a single endogenous variable ($n = 1$) and if the observation of y_{t-1} is not available but the observation of y_{t-2} is, the argument can unfold from the equation $y_t = b^2 y_{t-2} + \varepsilon_t + b\varepsilon_{t-1}$.) We shall assume implicitly in what follows that this sort of substitution, concerning the $x_{i,t-\tau}$ and $y_{j,t-\tau}$ that are still unobserved at time t, has been made and that the term ε_t gathers together all the uncertainties which affect the use of the model at the time of the decision.

To assume that all the "predetermined" variables are known is obviously the same as assuming that, while the decider does not know the current errors ε_t, he or she does know the past errors $\varepsilon_{t-\theta}$ for $\theta > 0$, since they can be calculated from system (186) rewritten after substituting $t - \theta$ for t. We shall take this into account in what follows.

(iii) The choice of the z_{kt} is determined so as to achieve, as well as possible, some objectives. We shall assume that these directly concern only endogenous variables y_{jt}. The assumption is not too restrictive: it does not imply that all the endogenous variables are included in the definition of the objectives (in practice, many endogenous variables are introduced for their explanatory power rather than for their intrinsic interest). Nor does it prevent some instruments z_{kt} for being included (we just have to insert the corresponding variables once in the list of instruments with their original labels, but also once in the list of endogenous variables with new labels; additional trivial equations will simply be needed in order to identify two corresponding labels).

The linear model displayed in this subsection differs only in appearance from those on which the mathematical theory of optimal control is constructed in mechanics. The latter would use, for example, the following more compact form instead of system (186):

$$\begin{aligned} u_t &= Au_{t-1} + Cz_t + \nu_t, \\ y_t &= Bu_t. \end{aligned} \tag{187}$$

This representation contains not only a new vector of non-observable errors ν_t but more importantly a new vector u_t made of observable variables, the so-called "state variables" u_{lt}. There are as many such variables as is necessary to get system (186) into the form of (187) thanks to appropriate definition of the vectors u_t, ν_t and the new matrices A, B and C. The state variables will so contain all those lagged values of the instruments which feature in

(186) with non zero coefficients, each couple (instrument, lag) corresponding to a distinct state variable. Among the state variables will also be found the exogenous variables on surrounding conditions, each one as many times as required for the proper introduction of lagged values. Similarly, predetermined endogenous variables will appear. Finally the error ε_{jt} will be captured in some way through the j component of Bu_t as resulting in particular from ν_t (the reader is advised to construct a form (187) for a simple example of a reduced form (186)).

To study the problems raised by the choice of macroeconomic stabilization policies, we can usually restrict ourselves to rather special cases of system (186). We shall do so in Section 3.7. But, it was useful to set out the linear system (186) in full generality, in order to appreciate the nature of restrictions contained in special cases and to have an intuitive understanding of possible generalizations of these cases. The first restriction, to be now introduced for pedagogical purpose, will be a major one because it will rule out the type of intertemporal dependence which permits stabilization. In the two following subsections we shall indeed discuss the choice of the optimal policy which would be appropriate if the model were purely static, that is, if the choice of the instruments z_{kt} in period t were totally independent of the choice of the $z_{k,t+\tau}$ in future periods $t+\tau$. We shall so become familiar with an approach which will be applied to the more general dynamic context in the following Section 3.7.

3.6.3. *Optimal decisions independently of time and uncertainty*

First of all, let us consider a standard general model for cases where neither time nor uncertainty appears. The objectives of policy concern some endogenous variables. These objectives are sometimes characterized as *targets*, that is as values to be reached by the variables in question. But such values are most often already the result of a decision, in which feasibility was taken into account. We shall therefore rather say that the objective is to meet, as well as possible, the preferences of the government in charge. In other words, we shall look for an "optimal decision"[194]. This formalization was often called the Frisch–Tinbergen theory of economic policy.

(i) A vector z with p numerical components z_k has to be chosen, each component defining the intensity of use of a particular instrument. This vector may be constrained to belong to a subset Z of R^p. A vector y of n endogenous

[194] About the evolution of ideas which, promoted in particular by R. Frisch, led to shift attention from targets to optimal decisions, see Section 7.1.8, pp. 392–394 in H. Theil, *Economic Forecasts and Policy* (North-Holland, Amsterdam, 1958).

numerical variables y_j characterizes the results of interest. A decision maker has preferences between the different values that this vector y can take[195]. It is not really restrictive so say that these preferences are expressed by a numerical objective function $U(Y)$ which the decider tries to maximize in the domain Y of R^n which defines the admissible values of y. Finally a system, here the economic system, determines how the choice of z influences the resulting vector y. We can express this dependence by $y = f(z)$, which may be a very compact notation. Notably, in many cases the function f depends on surrounding conditions, which may be represented by a vector x: a more explicit formalization would write $y = f(x, z)$, it being understood that the vector x would be exogenous with respect to the decision to be taken. Be that as it may, the problem of the decision maker is to maximise $U[f(z)]$ over Z subject to the constraint $f(z) \in Y$.

It is common to say that any particular vector y^0 is *controllable* if there is a value of the instrument vector z which allows y^0 to be achieved, that is, if y^0 belongs to $f(Z)$. We can say that the entire system represented by $f(z)$ is controllable if every y in Y is, that is, if $f(Z)$ covers the whole set Y. If this is the case the decision maker's problem can be decomposed. He can first look for a vector y^* which maximizes $U(y)$ on Y, then find a vector z^* such that $y^* = f(z^*)$. It is intuitively clear that a system is not controllable if the number of instruments is not rich enough with respect to the number of endogenous variables which appear in the formal representation of the system. Thus, for example, the instruments of macroeconomic policy, which are few, may be able to control the economic system if it is meant to contain only few endogenous variables of interest (for which reduced equations can be written). But too many variables would have to be introduced if a detailed control of the economy were aimed at, and then macroeconomic policy would be unlikely to suffice.

(ii) This standard model has often been chosen to present the logic and study of economic policy. It assumes that the objective function $U(y)$ is made explicit, even if only roughly. This is perfectly natural, since no economic policy recommendation could be formulated if we were unable to know what the objectives were and what weight should each of them receive when they cannot all be fully achieved simultaneously. But we must recognize that such objective

[195] In some applications the variables of interest are also assumed to directly include some instruments. This then may reflect other institutional constraints on the extent to which these instruments can be used, other than those incorporated in the definition of the set Z. Or it may reflect the desirability of relatively smooth changes in instruments. In such cases the objective function ought to written here as $U(y, z)$. For simplicity we shall not take this possible feature into account in this section and the next one.

functions are rarely explained. A government will rarely admit being ready to accept a certain increase in inflation or a certain increase in the foreign deficit in order to reduce unemployment in a given proportion. Neither will often a government admit being ready to accept the opposite. This is one of the reasons why advising in the choice of economic policy is so delicate.

Faced with this state of affairs, advisers have no other solution than to discover what the decision maker prefers by analysing, in particular, what her choices have been in various circumstances. Just as the set $f(Z)$, which could have been written $f(x, Z)$, depends on the surrounding conditions x, the vector z^* chosen also depends on the them. Each couple $[z^*, f(Z)]$ which has been observed gives information about the objective function $U(y)$, since $f(z^*)$ is the vector of $f(Z)$ which is optimal according to $U(y)$. It can be said that the decider reveals her preferences through her choices, which vary when the surrounding conditions vary, but always maximize the same function U. Determining the preferences so revealed by the decisions observed is rarely formalized[196]; but, as a mental operation, it always plays a role in economic policy advising.

(iii) The notions above can usefully be applied to *the linear-quadratic case*, that is to the particular case in which the function f is linear (with respect to the instrument variables) and the function U is quadratic (with respect to the endogenous variables) and in which the sets Y and Z respectively correspond to the whole spaces R^n and R^p. System (186), which for the moment would contain neither stochastic errors nor lagged variables, would give exactly the expression for the function f if this function was also linear in the variables x_i on surrounding conditions. Here, we simply write:

$$y = Cz + a, \tag{188}$$

where the vector a will be assumed to depend on the x_i in way which has not to be made precise.

Assuming a quadratic objective function and giving it a convenient form is natural, at least as a first approximation, on the basis of the following considerations. Two categories can be made among the endogenous variables y_j occurring in the representation of the economic system, as we had already occasions to see. The "variables of interest" directly matter for judging the outcome of the policy; they are naturally arguments of the objective function. The other

[196] See, however, C. van Eijk and J. Sandee, Quantitative determination of an optimum economic policy, *Econometrica* (January 1959). For references to various studies in the same vein and for a presentation of a method for determining implicit preferences, see L. Hordijk, A. Mastenbrock and J. Paelink, Contributions récentes à l'étude empirique de fonctions de préférences collectives, *Revue d'économie politique* (July–August 1976).

endogenous variables were introduced in the modellization of the phenomena but have not to feature in the objective function. There are usually many such endogenous variables in the structural form of macroeconometric models. But in this part of our discussion we find it convenient to work on reduced forms, such as system (186). Such being the case, we have nothing to lose (and much to gain in simplicity) in discarding all reduced equations except those in which the left-hand member is a variable of interest. This is why we now interpret the vector y in (188) as containing the endogenous variables of interest but no other. All the components of y then appear in the objective function.

Moreover, it is common to consider that for each variable of interest y_j there is a particular value y_j^* which would be the most satisfactory if it could be achieved with no sacrifice for the values reached by the other variables of interest. This value may in practice be out of reach, but it may be taken as the reference for y_j in the objective function. If moreover this function is to be quadratic, it is naturally defined as:

$$U = -(y - y^*)' W (y - y^*), \tag{189}$$

where W is a symmetric positive definite matrix. This matrix may in particular be diagonal, in which case the loss function $-U$ would be a weighted sum of squared discrepancies $(y_j - y_j^*)$; but we do not make that assumption.

Since the instrument vector z may be chosen with no constraint ($Z = R^p$), the set $f(Z)$ of the vectors y which are controllable is an hyperplane in $R^n = Y$, n being now the number of endogenous variables appearing in y, both in the reduced form (188) and in the objective function (189). The dimension of the hyperplane is then equal to the rank r of the $(n \times p)$ matrix C. If $r = n$, the normal case when there are so many instruments that $p \geqslant n$, any vector of R^n is controllable, hence in particular $y = y^*$, which obviously maximizes U.

Let us consider the more interesting case in which $r = p < n$: the number of instruments is smaller than the number of variables of interest. Substituting (188) for y in (189) we find the following function of z:

$$U = -z'C'WCz + 2(y^* - a)'WCz - (a - y^*)'W(a - y^*). \tag{190}$$

The first order condition for z to maximize this function is:

$$z = z^* = (C'WC)^{-1} C'W(y^* - a) \tag{191}$$

and this gives indeed the maximizing vector of U, leading to:

$$U^* = -(a - y^*)' [W - WC(C'WC)^{-1} C'W] (a - y^*). \tag{192}$$

Except in the special cases when x would be such that $a - y^*$ lead to $U^* = 0$, the maximum is negative. (For the proof, see the theory of quadratic forms and take account of $p < n$.)

3.6.4. *Certainty equivalence*

Let us now consider the case where non-observable stochastic shocks affect the dependence between the instruments and the endogenous variables. Let these shocks be represented by a vector e. Hence, the economic system is formalized by the equation $y = f(z, e)$. The value achieved by the objective function $U[f(z, e)]$, depends not only on the values chosen for the decision variables but also on the values assumed by the random vector e. As in other parts of this book and for the same reasons, we shall focus on the decision rule maximizing the mathematical expectation of U. We thus come back to a problem which is formally similar to the one discussed in the foregoing subsection where uncertainty was absent.

(i) It is interesting to consider the linear-quadratic case in which not only the function U is quadratic and the function f linear, but also the random vector appears as a purely additive term. Equation (189) applies whereas (188) is replaced by:

$$y = Cz + a + e. \tag{193}$$

Moreover in this linear-quadratic case the probability distribution of the vector e is assumed to be independent of z and such that:

$$E(e) = 0, \qquad E(ee') = V. \tag{194}$$

Hence, we can easily calculate the expression for the mathematical expectation of U, and we find:

$$\begin{aligned} EU &= -z'C'WCz + 2\big(y^* - a\big)'WCz \\ &\quad - \big(a - y^*\big)W\big(a - y^*\big) - tr(VW), \end{aligned} \tag{195}$$

where $tr(VW)$ is the trace of the matrix VW, that is, the sum of its diagonal elements. This result is remarkable, since it is close to be identical to Equation (190) holding in the absence of uncertainty: the only change is the subtraction of the constant $tr(VW)$. Obviously this change has no effect on the optimal choice of vector z: the optimal policy remains the same as when we ignored the presence of random shocks.

We say that, "*in the linear-quadratic case, the vector e of random shocks can be replaced by its mathematical expectation, which constitutes its certainty* *equivalent*". We can indeed substitute for the stochastic model the deterministic model (188), which is equivalent for the choice of the optimal policy. This is a very interesting property since it means that, in policy analysis, we might be justified in ignoring the presence of random disturbances and thus give a justification for a current practice.

(ii) But this convenient property only applies rigorously to the linear-quadratic case. We can easily see this if we have another look at the standard model in which the functions f and U are just assumed to have derivatives of all orders. Let otherwise take the simplest case: e is a random numerical variable with, in addition, $n = p = 1$ (a single instrument and a single endogenous variable).

Writing $G(z, e)$ for the function $U[f(z, e)]$, note that maximization of EU requires:

$$EG'_z(z, e) = 0. \tag{196}$$

This equality is not only necessary but also sufficient if EG is a concave function of z, a property that we assume. If we replace e by its zero mathematical expectation we can define the value $\hat{z}$ of z which maximizes $G(z, 0)$ and hence satisfies:

$$G'_z(\hat{z}, 0) = 0. \tag{197}$$

To study the possible difference between $\hat{z}$ and the solution z^* of (196), we can develop, up to second order terms, the function $G'_z(z, e)$ in the neighbourhood of $(\hat{z}, 0)$, take into account (197) and $E(e) = 0$, write $E(e^2) = \sigma^2$ and in this way replace (196) by the approximation:

$$\big(z - \hat{z}\big)G''_{z^2} + \frac{\sigma^2}{2}G'''_{ze^2} = 0. \tag{198}$$

The derivatives are evaluated at the point $(\hat{z}, 0)$ and we ignore the second order term in $z - \hat{z}$. Equation (198) implies the following approximation for z^*:

$$z^* = \hat{z} - \frac{\sigma^2 G'''_{ze^2}}{2G''_{z^2}}. \tag{199}$$

Note that the fraction evaluating the opposite of the corrective term is indeed equal to zero in the linear-quadratic case because its numerator is then equal to zero. The concavity of G means that the denominator is negative. The corrective term thus has the same sign as the third derivative G'''_{ze^2}, which is obviously not equal to zero in general. The mathematical expectation of the error e cannot in general play the role of an exact certainty equivalent in policy analysis.

It is intuitively clear that the difference between z^* and $\hat{z}$ comes from the fact that the optimal decision reflects a certain risk-aversion. If for example $y = f(x, e) = x + e$, then $G''_{ze^2} = U'''$. The optimal value z^* is larger than $\hat{z}$ if the curvature of the objective function is lower to the right of its maximum than to the left: for protection against the risk of a low value of y and thus of $U(y)$ the decision maker must choose a value of z which leads to a slightly higher y than the value which maximizes U.

(iii) However, Equation (199) draws our attention to the fact that $z^* - \hat{z}$ is proportional, as a first approximation, not to the standard deviation of the error e but to its variance σ^2. If the probability distribution of e has a small standard deviation σ, the correction $z^* - \hat{z}$ is second-order small in σ. It may then be taken as negligible in some applications.

Such a property holds in general. Given differentiability and regularity conditions, which would be long to describe here, it can be shown that, up to first order terms, the optimal decision is the same as the one that would be computed if the stochastic terms were replaced everywhere by their mathematical expectations. To characterize briefly this property we say that "*the mathematical expectation constitutes a certainty equivalent to the first order*".

The properties of certainty equivalence, which hold exactly in the linear-quadratic case and up to the first order in the more general regular case, are not specific to static models. They also hold for the problem studied in the next section, that is, for sequential decisions taken within the framework of dynamic models. Explaining why they do so would require introducing a rather cumbersome formal structure[197]. Let us simply record that these properties justify, to some extent, simplifications we are going to make in the next section.

3.6.5. *The limits of certainty equivalence properties*

It would, however, be a dramatic misinterpretation to believe that the properties in question allow, in macroeconomic policy making, a full neglect of the random features of the shocks to the economy.

(i) This would indeed, in the first place, contradict the results examined in Section 7.6 of Chapter 8, where we saw that, depending on the relative importances of shocks affecting respectively the demand for goods and the demand for money, it was preferable for monetary authorities to control the rate of interest or the quantity of money. The reason clearly appears if we compare the

[197] See H. Theil, *Optimal Decision Rules for Government and Industry* (North-Holland, Amsterdam, 1964), and E. Malinvaud, First order certainty equivalence, *Econometrica* (October 1969).

treatment discussed in that section with the one just made. The certainty equivalence properties concern the relationship between a given list of instrument variables and a given list of endogenous variables. In Section 7.6 of Chapter 8 we were wondering about the choice of a variable to be controlled, that is, the choice of an instrument variable in our present terminology[198]: if it was the interest rate, then the quantity of money was endogenously determined and vice versa.

(ii) The properties do not apply beyond the regular case, which excludes cases where the first derivatives of functions appearing in the economic system exhibit discontinuities, *a fortiori* cases allowing regime switches such as those discussed in Part 4 of Chapter 7 and at various places in this chapter, particularly in Section 4.1 below. Even when the properties formally apply, the certainty equivalent system may provide a poor reference for forecasting and policy analysis if the probability distribution of the disturbances has a substantial dispersion, particularly when important non-linearities appear.

Since these two features, large dispersion of error terms and important non-linearities, concern about all structural models used in practice, the conclusion follows: we may worry about the relevance of the point estimates of forecast values and multipliers which are obtained from the so-called certainty-equivalent system. In order to calm down such worries the proper solution consists in giving probabilistic assessments of forecasts and multipliers.

Econometric methodology worked out analytical procedures suited for such assessments, but then only for standard linear cases, which are essentially much simpler than actual structural models. To deal with the latter there is no alternative to "stochastic simulation". Even if the deterministic evaluation of forecasts and multipliers from the certainty-equivalent system is retained, stochastic simulation otherwise remains necessary in order to estimate the uncertainty of the evaluation.

(iii) A *stochastic simulation* is obtained by what was often called "the Monte Carlo method". Artificial pseudo-random numbers are drawn in such a way as to mimic what would have been a random draw from the probability distribution of the vector of disturbances, ε_t say. The model is then solved, for given values of the exogenous variables x_t and z_t, a solution which gives the vector y_t of the corresponding values of the endogenous variables. The operation

[198] We know that an instrument variable, which a macroeconomic model assumes to be controlled, within years or quarters, say, may much differ from the instrument actually used day after day by technicians in charge of implementing the policy. For instance, a control of the interest rate may be achieved by daily open market operations, or a control of the quantity of money by the daily choice of the interest paid on Treasury bills.

is repeated many times, for the same values of x_t and z_t, and for other relevant sets of values as well. Thus, an empirical distribution of the endogenous vector y_t is obtained, as well as its changes as a function of some changes in x_t and z_t. Measures of the dispersion of this distribution may be taken as assessments of the uncertainty that surrounds the model's forecasts.

Actually, stochastic simulations have to also take account of a more indirect source of uncertainty, the estimation error in the model's coefficients, which may substantially affect the uncertainty of forecasts and is the main reason for uncertainty of multipliers. But, applied to the estimation stage, the Monte Carlo method also provides measures of this second source of uncertainty.

We shall not examine here the details in the production of stochastic simulations. For these the reader may refer to R. Fair (1984, op. cit., Chapter 9) and to K. Wallis (1995, op. cit., Part 4), which also reports a number of interesting results. We shall simply comment on what seems to have been learned about the bias in the deterministic forecasts obtained from the certainty-equivalent system. On the one hand, the qualitative conclusion drawn from stochastic simulations has often been that this bias, due to the non-linearity of structural models, is of little practical importance. On the other hand, whether the bias matters or not is likely to very much depend on the particular economic context. For instance, P. Fisher and M. Salmon[199] found that, in dynamic simulation of multi-period forecasting, the bias can build up as the forecasting horizon is extended farther into the future; this is particularly likely when the solution is moving away from its initial conditions.

3.7. Optimal control in macroeconomic policy-making

The dynamic framework introduced in Subsection 2 of the preceding section should now be applied, together with an objective function and a certainty equivalence property generalizing those exposed in Subsections 3.6.3 and 3.6.4. This would result in an optimal control methodology which was set up mainly by G. Chow for application of structural models to the design of macroeconomic policy[200]. We shall, however, refrain from trying and exposing the mathematics of the approach somewhat systematically. Not only would such an exposition be heavy. But also experience and reflection about the possibility and advisability of actual use, for policy making, of optimal control on

199 P. Fisher and M. Salmon, On evaluating the importance of nonlinearity in large macroeconometric models, *International Economic Review* **27** (1986) 625–646.

200 G. Chow, *Analysis and Control of Dynamic Economic Systems* (Wiley, New York, 1975); G. Chow, *Econometric Analysis by Control Methods* (Wiley, New York, 1981).

structural models showed that such possibility and advisability were limited. The lasting contribution of the approach is rather to have led economists who worked on policy into the habit of thinking about policy as part of a system through a reaction function and a feedback rule.

The idea of optimal control as a technique in applied macroeconomics arose at the time marking a climax of the confidence in structural models. But during the late 1970s and early 1980s this confidence was hard hit. After a period of infatuation for "models" and for the supposed power of economic policies based on them, opinion in the general public strongly reacted when business conditions turned out to become bad; disappointment and dissatisfaction led to excessive distrust. Although less extreme in their judgment, both in 1970 and in 1985, most economists themselves were subject to the same wave of optimism and pessimism. Specialists knew that the forecasting performance of structural models, as we surveyed them in Section 3.4, were not high and that valid criticism affecting the supposed accuracy of models had been raised by a careful discussion of the methodology, a discussion which will be further referred to. Since the models did not seem to provide such a good representation of the economic system, could they really serve in the same way as some physical models were serving for optimally controlling the trajectory of a spacecraft?

Analogy with engineering systems was still more shaken by acceptation of the idea that the economic models should incorporate expectation variables treated in a rational and model-consistent manner. After the Lucas' critique, which argued that private agents were forming expectations about future policies, and suggested that rational expectations of these policies should be assumed in appropriate models, F. Kydland and E. Prescott introduced the credibility issue and concluded: "There is no way control theory can be made applicable to economic theory when expectations are rational"[201]. If so, a careful treatment of optimal control has relevance only for those structural models which do not accept the Lucas' critique and rather assume expectations to be adaptive. Again, this narrows the scope of cases to which would apply the methodology developed for the optimal control of dynamic economic systems[202].

However, it would not be wise here to fully neglect this methodology. Indeed, it provides a correct framework for the treatment of some cases and for an approach to a number of side problems which often occur in macroeco-

[201] F. Kydland and E. Prescott, Rules rather than discretion: the inconsistency of optimal plans, *Journal of Political Economy* **85** (1977) 473–491.

[202] Discussion of the relevance of optimal control for the elaboration of economic policy went up to parliamentary level in the UK. Historians of our discipline will like to study R. Ball, *Report of the Committee on Policy Optimisation*, her Majesty's Stationary Office, London (1978).

nomic policy analysis. Our ambition here is accordingly limited: neither full exposition nor full neglect.

For a complete theory of the logic underlying the formulation of countercyclical policy under adaptive expectations, it is essential to include four aspects of the context within which this policy takes effect:

- time and the many different lags in adaptations and reactions,
- uncertainty as to knowledge of the economy and to the forecast of changes in the economic environment,
- the structure of the information which is progressively reaching policy makers,
- the non-linear character of numerous phenomena.

If developed even for a more limited scope than was envisaged in the 1970s, this theory would be formidably complex. For our present much more modest purpose we shall just attempt at understanding the logical nature of certain fairly natural rules that could be adopted for a stabilization policy. After having seen how a general argument can claim to show that a totally predetermined policy has little chance of being optimal, we shall examine, precisely but in very simple cases, how can be defined optimal sequential rules, according to which the policy would be progressively determined on the basis of information received. We shall then base our arguments on linear models and show how the values to be given to the instruments ought to adapt to changes in the exogenous variables on surrounding conditions. We shall also see how the policy ought to deal with the tradeoffs between objectives which are not fully compatible with each other. We shall end with the examination of an application of the approach.

3.7.1. *An argument for contingent rules*

Determining a policy rule consists in fixing in advance the calculations which will be made in order to give values to the instruments. Assume that the rule must be applied up to an horizon including T successive decision dates ($t = 1, 2, \dots, T$), and denote in a general way by I_t the information which will be available at date t when the value of the corresponding instrument vector z_t will have to be chosen. A policy rule is a set of T functions g_t to be used to calculate the vectors z_t:

$$z_t = g_t(I_t), \quad t = 1, 2, \dots, T. \tag{200}$$

Defining a rule consists in specifying, right from date 1, the set of the T functions g_t.

In general, the information I_t contains not only the initial information I_1 and the values taken by exogenous variables on surrounding conditions,

$x_1, x_2, \ldots, x_t$, but also the history of the values previously chosen for the instruments, $z_1, z_2, \ldots, z_{t-1}$, as well as observation of past decisions results on the values of endogenous variables, $y_1, y_2, \ldots, y_{t-1}$. In general the decision z_t is sensitive to each one of these elements: for example, g_t implies that z_t changes in terms of what y_{t-1} has been. We say that the rule includes feedbacks, that it is "conditional", "adaptive" or, more specifically in the decision literature, "sequential". (The literature on servomechanisms has led some authors to use the expression: "a closed-loop rule".)

It is conceivable, however, to decide in advance on the whole sequence of values to give to the instruments and not to allow current information to have any influence. Fixing once and for all the growth rate of the money supply corresponds to a rule of this type. Formally, this amounts to considering special cases of Equations (200) and calculating the z_t by equations such as:

$$z_t = h_t(I_1), \quad t = 1, 2, \ldots, T. \tag{201}$$

The policy rule is thus "fixed" or "unconditional" (or "open-loop").

Since information I_t includes information I_1, it is clear that (201) is a special case of (200), at least if the choice of functions g_t is left completely free. Even if, for reasons of simplicity, we limit our attention to particular classes of functions g_t and h_t, it will still be likely in practice that (200) includes (201) as a particular case. In these conditions, the principle of choosing the optimal fixed rule could not be superior to the principle of adopting the optimal sequential rule. Moreover, it seems unlikely that the optimal sequential rule could ever be found to be a fixed rule. In the simple cases we shall examine below, such a possibility does not arise; *a fortiori* it should not be found in more complex cases. Thus, we have good reasons to say that searching for the optimal sequential rule is a better strategy than choosing a fixed rule. How is it then that quite a few economists are recommending fixed rules? Various types of argument can be put forward[203].

Most of them stem from distrust of those responsible for economic policy. This distrust may concern different stages in the conception and application of policy. On may fear, first, that politicians give too much weight to the short term compared with the long term; thus the choice of their objective function is questioned and the idea is accepted that long-term objectives would be fairly well obtained by the use of a proper fixed rule, the lack of short-term adaptations being then viewed as less disturbing than the neglect of the long run. One

[203] For a more precise study of the link between the above immediate argument and certain discussions about the scope for active economic policies, the reader can refer to W. Buiter, The Superiority of contingent rules over fixed rules in models with rational expectations, *Economic Journal* (September 1981).

may also consider that knowledge of the economy is too uncertain and that active policies have not been enough studied to justify the claim that a truly optimal rule has been found; according to such a view, the policy would stand on firmer ground if applying an extremely simple, hence fixed, rule. Finally, one may consider that the actual ways in which decisions are put into practice rule out sequential control by public authorities; government services would have a too complex organization and be too accustomed to routine-work to execute rules other than simple ones, which would have to be fixed once and for all.

More sophisticated arguments have been put forward, which take into account visions of what rational expectations of private agents ought to be. Aiming at stabilization, the contingent rules usually considered would moderately stimulate activity in depression times at the cost of minor effects on inflation. But, if applied, such a rule would generate for government the incentive to exceptionally deviate by an unexpected larger stimulus which, since unexpected, would not much increase inflation. Rational agents would realize the existence of the incentive; they would consider permanent adherence to the contingent rule incredible; they would accordingly have higher price expectations than assumed in the definition of the contingent rule. Fixed rules would be more immune from such lack of credibility.

We shall not look more deeply now into this type of argument, which was discussed within a simple framework in Part 3 of Chapter 8. It clearly points to a difficulty in the conception of the optimal control approach, but a difficulty which would be taking us outside the adaptive expectations framework accepted in this section.

3.7.2. *An optimal sequential strategy*

To see how an optimal rule can be determined, consider, to start with, the situation where there is a single objective variable, where y_t is thus a number. In order to simplify matters we shall also keep to the case of a single exogenous variable on surrounding conditions x_t and a single instrument z_t.

For example, let y_t be the level of aggregate demand, z_t the level of government demand, and x_t firms' intentions about their investments as known from a business survey. The objective is to reach a target level for aggregate demand which would allow satisfactory values of employment and the inflation rate. The level of aggregate demand is thought to be a linear function of present and past values of government demand as well as firms' investment intentions; in addition, this function includes a random perturbation ε_t which cannot be controlled and is observed only after the fact.

Such a model looks familiar to us; indeed, it is deliberately chosen simple. The idea that the instrument is government demand suggests fiscal policy and the theory of the multiplier. But we could just as well envisage z_t as the instrument of monetary policy and refer to the theory exposed in Section 2.3 of Chapter 7, keeping the same interpretations of y_t and x_t.

Be that as it may, Equation (186) now corresponds to a single equation. To rewrite it let us once again use the lag operator L (by definition $Lx_t = x_{t-1}$). This leads to the following expression for the law of motion of the economic system[204]:

$$y_t = A(L)x_t + B(L)y_t + C(L)z_t + \varepsilon_t \tag{202}$$

with by definition:

$$A(L) = \sum_{\tau=0}^{h} a_\tau L^\tau, \qquad B(L) = \sum_{\tau=1}^{h} b_\tau L^\tau,$$
$$C(L) = \sum_{\tau=0}^{h} c_\tau L^\tau. \tag{203}$$

The target for economic policy is a well defined value $\hat{y}$ of y_t. We assume it does not change over time. How should the successive z_t be chosen so as to lead to values of y_t as close as possible to the target $\hat{y}$?

(i) Before answering the question we must specify the information held by the decision-maker and the stochastic properties of the process of perturbations ε_t. We assume that, before choosing the value of z_t, the decider knows the $x_{t-\tau}$ for $\tau = 0, 1, \ldots$, the $y_{t-\tau}$ and the $z_{t-\tau}$ for $\tau = 1, 2, \ldots$. We also assume that the mathematical expectation of ε_t is zero and its variance is constant:

$$E(\varepsilon_t) = 0, \qquad E(\varepsilon_t^2) = \sigma^2. \tag{204}$$

In addition, we assume to start with that the successive ε_t are independent of each other.

The certainty-equivalence property applies here rigorously if the objective function is proportional to *minus* the square of $y_t - \hat{y}$; more generally it applies up to the first order if the function is equal to zero for $y_t = \hat{y}$, is negative otherwise and has derivatives of order 1 and 2. So the optimal values of the z_t are just as if the ε_t were zero:

$$C(L)z_t = \hat{y} - A(L)x_t - B(L)y_t. \tag{205}$$

204 The linear equation has no constant term. This makes writing the following equations simpler, and has no substantial importance thereafter.

This equation expresses how the value of z_t has to be calculated in terms of the available information I_t at time t: this optimal value adapts to the observations made of x_t and of the earlier[205] values $x_{t-\tau}$ and $y_{t-\tau}$.

The equation defines a *sequential decision rule* or a *reaction function*. As a result in this case $y_t = \hat{y} + \varepsilon_t$: the endogenous variable oscillates randomly around its target. The optimal policy thwarts all causes of deviation from the target except the uncertainty which remains at the time when the value of the instrument is chosen. This is not a deep property since it rather directly follows from the assumptions we introduced; nevertheless, it is suggestive. It is a direct translation of the fact that, in the certainty-equivalent model, the target $\hat{y}$ is controllable in every period.

What will then be the path of the instrument variable z_t? Here as well, the answer is easy since (205) implies:

$$z_t = \frac{-A(L)}{C(L)} x_t + \frac{1 - B(L)}{C(L)} \hat{y} - \frac{B(L)}{C(L)} \varepsilon_t. \tag{206}$$

The optimal value of instrument z_t will follow a path which depends on:
- the history of values taken by the exogenous variable x_t,
- the nature of the dynamic dependence featured in system (202).

(ii) Even though the optimal policy will result in stabilizing the endogenous variable of interest, except for random errors, this does not mean that stabilization will apply to the instrument. The values given to z_t will have to change as much as required to compensate for changes in the current and past values of the exogenous variable x_t and in the past values of the shocks ε_t. The path of z_t will then have to fluctuate, even much so in some cases.

We can easily see it in considering for instance the very simple case where there would be no uncertainty ($\varepsilon_t = 0$) and the law of motion would result from:

$$A(L) = a_0, \qquad B(L) = 0, \qquad C(L) = c_0 + c_1 L. \tag{207}$$

Equation (205) would reduce to:

$$c_0 z_t + c_1 z_{t-1} = \hat{y} - a_0 x_t. \tag{208}$$

[205] More precisely the definitions of $A(L)$ and $B(L)$ by (203) show that $B(L)y_t$ does not depend on the current value of y_t and that all lagged values of x_t and y_t have to be known starting from $t = 1 - h$. The initial information I_1 is assumed to be such that this condition is met from the beginning date $t = 1$.

Even with a constant value $x_t = \hat{x}$ for $t = 1, 2, \ldots$, it would not be true in general that z_t would be constant and equal to the value $\hat{z}$ given by $(c_0 + c_1)\hat{z} = \hat{y} - a_0\hat{x}$. It would still be a function of the initial value of z_0:

$$z_t - \hat{z} = \left(\frac{-c_1}{c_0}\right)^t (z_0 - \hat{z}). \tag{209}$$

In the case where c_1 would be lower than c_0 in absolute value, z_t would tend to the constant $\hat{z}$. But in the opposite case, z_t would deviate more and more from $\hat{z}$. For example, if $c_1 > c_0 > 0$, the values given by (209) alternates with an increasing amplitude around $\hat{z}$. In order to maintain y_t at the target value $\hat{y}$ it is necessary, in each period, to cancel the effect of the deviation $z_{t-1} - \hat{z}$ of the previous period and this requires an even larger deviation in the opposite direction.

Such an alternating and explosive path could obviously not be maintained indefinitely since the linearity of the model would appear unacceptable in practice for very large changes; nevertheless the example shows that *stabilization of the variables of interest does not logically imply stability of the instrument.* It is possible to imagine that a *stop and* go fiscal policy might be appropriate for stabilizing the level of aggregate demand; the same could be true for a monetary policy which could alternate periods when credit would be scarce and expensive with periods when it would be plentiful and cheap.

In general Equation (206) shows that large fluctuations, to be imposed to the instruments by an optimal policy rule, can be explained by two causes: either by the instability of the values of exogenous variables on surrounding conditions, or by the lag structure of the relations through which the instruments affect the endogenous variables of interest.

(iii) We have assumed that successive perturbations were independent of each other; but this is not an essential assumption. It is easy to now see how the previous analysis can be generalized. Assume for example that $\{\varepsilon_t\}$ follows a known moving-average stationary stochastic process. Let us write:

$$\varepsilon_t = \big[1 + D(L)\big]\eta_t \quad \text{with} \quad D(L) = \sum_{\tau=1}^{h} d_\tau L^\tau, \tag{210}$$

where η_t are independent identically distributed random variables, with zero mean.

Since the $\varepsilon_{t-\tau}$ are known for $\tau > 0$ at the time when the value of z_t is chosen, so also are the $\eta_{t-\tau}$, which can be recursively computed from the $\varepsilon_{t-\tau}$. Consequently, to apply the certainty equivalence property, we should no longer replace ε_t in the law of motion by its unconditional mathematical expectation, which is zero, but by its conditional expectation, information available in period t being taken into account[206]. The perturbation ε_t should thus be replaced by $D(L)\eta_t$, which is its mathematical expectation conditional on knowledge of all $\eta_{t-\tau}$ for $\tau > 0$.

The optimal rule no longer is given by (205) but rather by:

$$C(L)z_t = \hat{y} - A(L)x_t - B(L)y_t - D(L)\eta_t. \tag{211}$$

The result then is that the endogenous variable will take the value $y_t = \hat{y} + \eta_t$. The optimal policy again thwarts all causes of deviation from the target except those coming from whatever uncertainty still remains at the time when the value of the instrument is chosen.

The same property of controllability often applies when there are several variables of interest and when the number of instruments is equal to the number of these variables. But complications may appear: for example, aggregate demand regulation aims at two objectives: high employment and a low rate of inflation; two instruments can be used: public spending and the interest rate, say; but, given that both these instruments act through one and the same variable, aggregate demand for goods and services, the policy cannot usually reach the two independent targets, for employment and inflation: these targets are not controllable, even if this adjective admits the deviations reflecting the existence of uncertainty. We are back to the situation studied in Chapter 8, Part 2.

To be aware of the control problems in the context of this sort of situation, and to know how they can be solved, we are now going to consider a case where precisely there are two target variables and a single instrument. Our line or argument will suggest the direction in which generalization could be found, but we shall not explore these here.

3.7.3. *An optimal strategy with tradeoff between objectives*

(i) Keeping a single exogenous variable on surrounding conditions x_t and a single instrument z_t, we assume that y_t and ε_t are vectors with two elements. We are now going back for simplicity to the case where the ε_t are mutually independent of each other. These random vectors have a zero mathematical expectation and the same covariance matrix:

$$E(\varepsilon_t) = 0, \qquad E\left(\varepsilon_t \varepsilon_t'\right) = \Omega. \tag{212}$$

Finally, for simplicity again we consider a very special case of the general linear dynamic model (185): lagged values of x_t and z_t do not appear whereas

[206] A. Duchan proved that this is the correct version of the certainty equivalence property when the random terms are serially correlated: A clarification and new proof of the certainty equivalence theorem, *International Economic Review* (February 1974).

there is a single lagged value for the endogenous vector y_t. Hence, the structural model can be written as follows:

$$y_t = By_{t-1} + ax_t + cz_t + \varepsilon_t, \tag{213}$$

where B is a square matrix of order 2, whereas a and c are vectors with two components each.

We are looking for a *sequential decision rule*, which will have the same form for all t. More precisely we are going to determine the optimal policy from amongst those which can be expressed in a linear form:

$$z_t = \alpha x_t + \beta_1 y_{1,t-1} + \beta_2 y_{2,t-1} + \gamma, \tag{214}$$

where α, β_1, β_2 and γ are numbers to be determined. Note that this equation seems to be a fairly natural transposition to this example of Equation (205) which we obtained in the preceding subsection. We could prove that, under fairly general conditions, the optimal policy does indeed have this linear form. Equations (213) and (214) show that the vector of endogenous variables resulting from the policy will be determined by:

$$y_t = Py_{t-1} + qx_t + r + \varepsilon_t \tag{215}$$

with:

$$P = B + c\beta', \qquad q = a + \alpha c, \qquad r = \gamma c. \tag{216}$$

All we have to do now is to choose α, β and γ optimally.

For this we want to minimize the intertemporal loss function:

$$V = \frac{1}{T}\sum_{t=1}^{T} E\big(y_t - \hat{y}\big)'\big(y_t - \hat{y}\big), \tag{217}$$

where $\hat{y}$ is a vector of target values for y_t and T is a large number, which we shall even assume to be infinite in what follows. Except for the sign and the presence of the mathematical expectation operator, this function has the form (189), the matrix W being there the identity matrix (by a suitable choice of units for y_{1t} and y_{2t} we can always manage to give these two variables the same weight, in the objective function; the only restrictive assumption concerning within period choices is therefore to assume that W is diagonal). However, it is noteworthy, and restrictive for intertemporal choices, that the formula assumes no discounting of future losses. The assumption is made here for simplicity again (with discounting, the analytical approach would have an additional dimension of complexity).

The optimal sequential rule must select that value of z_t which achieves *the best possible tradeoff between two incompatible targets* $\hat{y}_1$ and $\hat{y}_2$. It obviously depends on what are the known properties of the sequence of the exogenous variables x_t. We have to choose a particular specification for this. One possibility for our present study might be to assume that this sequence is the realization of a stationary stochastic process. Almost equivalently we take this sequence as deterministic, but we assume that it has the same asymptotic properties as the sequence of observations generated by such a process. Moreover we restrict this reference process to being a pure white noise. In other words, we assume that the following limits hold when T increases up to infinity:

$$\frac{1}{T}\sum_{t=1}^{T} x_t \to \bar{x}, \qquad \frac{1}{T}\sum_{t=1}^{T}\left(x_t - \bar{x}\right)^2 = \sigma_x^2,$$

$$\frac{1}{T}\sum_{t=1}^{T}\left(x_{t+\theta} - \bar{x}\right)\left(x_t - \bar{x}\right) \to 0, \quad \text{if } \theta > 0.$$

(ii) Equation (215) can be written as:

$$y_t - \hat{y} = P\left(y_{t-1} - \hat{y}\right) + q\left(x_t - \bar{x}\right) + (I - P)v + \varepsilon_t \tag{218}$$

with:

$$v = (I - P)^{-1}\left[a\bar{x} + \left(\gamma + \alpha\bar{x}\right)c\right] - \hat{y}. \tag{219}$$

The limiting expression of the loss function, when T tends towards infinity, is obtained by calculations which we do not spell out here. They are conveniently made thanks to the moving average representation of the vector $y_t - \hat{y} - v$ ruled by the autoregression (218). According to this representation $y_t - \hat{y} - v$ is a linear function of current and past values of $x_t - \bar{x}$ and ε_t. The end result of the calculation is:

$$V = tr\left[\sigma_x^2 q q' + \Omega\left(I - P'P\right)^{-1}\right] + v'v, \tag{220}$$

where $tr\,X$ denotes, as usual, the sum of the diagonal elements of the matrix X.

As P and q do not depend on γ, the latter should be chosen so as to minimize $v'v$. Now the differential of v is given by:

$$\mathrm{d}v = (1 - P)^{-1}c\left[\mathrm{d}\gamma + \bar{x}\,\mathrm{d}\alpha + \mathrm{d}\beta'(1 - P)^{-1}\right] \tag{221}$$

(using here the formula $\mathrm{d}X^{-1} = -X^{-1}\,\mathrm{d}X \cdot X^{-1}$). Minimizing $v'v$ with respect to γ implies that the coefficient of $\mathrm{d}\gamma$ in $v'\,\mathrm{d}v$ is equal to zero, hence:

$$v'(1 - P)^{-1}c = 0, \tag{222}$$

which implies moreover $v'\,\mathrm{d}v = 0$.

Taking Equation (222) into account, the following result is obtained from calculations which we do not spell out here:

$$\begin{aligned} \mathrm{d}V &= q'\big(1 - P'P\big)^{-1} c\,\mathrm{d}\alpha \\ &\quad + c'P\big(1 - P'P\big)^{-1}\big(\sigma_x^2 qq' + \Omega\big)\big(I - P'P\big)^{-1}\,\mathrm{d}\beta. \end{aligned}$$

For the coefficients of dα and of dβ to equal zero we just need to impose:

$$q'\big(1 - P'P\big)^{-1} c = 0, \tag{223}$$

$$c'P = 0. \tag{224}$$

Equations (222), (223) and (224) determine the coefficient α, β and γ of the optimal sequential rule. In particular, the last equation implies:

$$\beta = \frac{-B'c}{c'c} \tag{225}$$

and consequently:

$$P = \left[I - \frac{cc'}{c'c}\right]B. \tag{226}$$

Let us go back to the comparison between Equation (213), which rules the evolution of the endogenous vector y_t in our model, and Equation (215), which shows its evolution as a result of the policy. We then see that matrix B, which appears in the dynamic behaviour of the economic system, is, after the adoption of the optimal sequential policy, multiplied by the matrix in the square bracket of (226). This latter matrix is singular since postmultiplied by c it leads to the zero vector. Its non-zero characteristic root is equal to 1, since subtracting I within the bracket and postmultiplying by any vector orthogonal to c again leads to the zero vector. Thus, P may be said to be "smaller" than B. The lagged vector y_{t-1} plays a smaller role in the determination of y_t. Let us moreover remember that our assumptions ruled out serial correlation of both x_t and ε_t. It may thus be said that, in this sense, the *optimal policy dampens the spontaneous dynamic path of the endogenous vector* y_t. It may even be said that there is complete dampening for the linear combination $c'y_t$ since (215) and (224) imply:

$$c'y_t = c'qx_t + c'r + c'\varepsilon_t. \tag{227}$$

Note also that the optimal rule is such that the influence of x_t on the linear combination $c'(1 - P'P)^{-1}y_t$ is eliminated, as (215) and (223) show. But influence of x_t on each of the components of y_t still remains. The optimal policy no longer fully thwarts all known causes of fluctuations of y_t away from its target, as it succeeded in doing in the preceding subsection.

(iii) We have to recognize that the expressions defining coefficients α, β_1, β_2 and γ of the optimal rule are hardly transparent. But interesting simplifications appear in the case where the matrix B is proportional to the identity matrix. Indeed, (226) then implies $Pc = 0$, hence also $(1 - P)^{-1}c = c$ and $(1 - P'P)^{-1}c = c$. Equations (222) and (223) reduce to:

$$v'c = 0, \qquad q'c = 0 \tag{228}$$

from which we calculate:

$$\alpha = \frac{-c'a}{c'c}, \qquad \gamma = \frac{c'\hat{y}}{c'c}. \tag{229}$$

So the optimal decision rule is written:

$$z_t = \frac{-c'}{c'c}\big(ax_t + By_{t-1} - \hat{y}\big). \tag{230}$$

Consequently, the process followed by y_t is:

$$y_t = \left[I - \frac{cc'}{c'c}\right]\big(ax_t + By_{t-1} - \hat{y}\big) + \hat{y} + \varepsilon_t. \tag{231}$$

Equation (227) is replaced by:

$$c'y_t = c'\hat{y} + c'\varepsilon_t. \tag{232}$$

Thus, the linear combination $c'y_t$ oscillates randomly around its target value $c'\hat{y}$.

To get a better understanding of the significance of these expressions in the case where B is equal to bI, write:

$$w_{jt} = a_j x_t + b y_{j,t-1} - \hat{y}_j, \quad j = 1, 2. \tag{233}$$

This variable w_{jt} measures, in a way, the difference $y_{jt} - \hat{y}_j$ which would be expected in the absence of the policy action on the instrument z_t. Equations (230) and (231) show that the optimal value of the instrument is a linear combination of the two variables w_{1t} and w_{2t}, just as is the resulting difference between y_{jt} and $\hat{y}_j + \varepsilon_{jt}$.

For example, if $c_1 = 1$ and $c_2 = 1$, we find

$$z_t = \frac{1}{2}(w_{2t} - w_{1t}), \qquad y_{jt} = \frac{1}{2}(w_{1t} + w_{2t}) + \hat{y}_j + \varepsilon_{jt}. \tag{234}$$

Imagine that y_{1t} is the inflation rate and y_{2t} is unemployment. Suppose the two values w_{1t} and w_{2t} are positive in period 1. Choosing a high value for z_t

would reduce unemployment but increase the rate of inflation. Since the two endogenous variables have the same weight in the loss function, the instrument is chosen so that the two expected deviations of each variable from its target have the same value, actually the mean of w_{1t} and w_{2t}. The optimal policy will reduce inflation if this tends to deviate more from its target than unemployment does, and vice versa.

These conclusions are not surprising. The interest of the study just made is to exhibit and spell out the logic behind them: the meaning of an optimal sequential decision rule, a precise mathematical procedure for the determination of the rule, some conditions under which calculus of the rule boils down to application of intuitively appealing formulas, etc. Even for the simple cases selected in this presentation, we had to carefully specify a number of points and the mathematics was not always trivial. We realize that dealing, at some level of generality, with the theory of optimal control of dynamic economic systems, would have to be quite technical.

3.7.4. *An application*

In order to see how the optimal control approach can be used for relevant econometric applications we may briefly refer to an example. We choose here that presented by J. Cooper and S. Fischer[207]. The model of the US economy is much simpler than any structural macroeconometric model ever built. It involves just four variables: two endogenous variables of interest and two policy instruments:

$$y_t = [u_t, g_t]', \qquad z_t = [a_t, m_t]', \tag{235}$$

where u_t in the unemployment rate, g_t the inflation rate, a_t the growth rate of the value of public expenditure on goods and services, m_t the growth rate of the quantity of money. A system such as (202) with $h = 2$ and x_t reduced to the constant dummy is fitted by a least-squares regression on quarterly data (if we want, we may say that the authors so obtained an approximation of an abridged reduced form of a structural model).

Assuming that, at the time when the values of the instruments a_t and m_t for quarter t are chosen, the previous values of all four variables have been observed, J. Cooper and S. Fischer look for a sequential linear policy rule. For the purpose they choose an objective function, which we shall not discuss here but involves target values $\hat{u}$ and $\hat{g}$ for u_t and g_t as well as a preferred value

[207] J. Cooper and S. Fischer, A method for stochastic control of nonlinear econometric models and an application, *Econometrica* (January 1975).

$\hat{m}$ for the monetary instrument. They so obtain equations giving the optimal values of a_t and m_t:

$$\begin{aligned} a_t &= 0.09(u_{t-1} - \hat{u}) + 0.10(u_{t-1} - u_{t-2}) - 0.06(g_{t-1} - \hat{g}) \\ &\quad - 0.16(g_{t-1} - g_{t-2}), \qquad (236) \\ m_t &= -0.45(m_{t-1} - \hat{m}) + 0.92(u_{t-1} - \hat{u}) + 1.22(u_{t-1} - u_{t-2}) \\ &\quad - 0.60(g_{t-1} - \hat{g}) - 1.82(g_{t-1} - g_{t-2}). \qquad (237) \end{aligned}$$

Note that the rule implies a much greater flexibility of monetary policy than of fiscal policy (but this is in part the result of the choice of the objective function), and also that these two instruments have to act almost in parallel, with a slight lead for m_t over a_t due to the presence of m_{t-1} in (237): parallelism is seen in the fact that the coefficients in (237) are almost exactly 10 times those in (236). This shows that, even though we have here two instruments for two variables of interest, the optimal policy is actually faced with a trade-off between unemployment and inflation, almost as in the case discussed in the preceding subsection. For this feature in particular, the result appears to be fairly classic.

Note also that the authors calculate the value that their loss function would have assumed over the sample period, depending on the policy rule applied: for the optimal rule given above (the result then was 49.6), for the policy actually implemented (58.5) and for a totally passive policy which would have maintained constant the monetary and fiscal instruments whatever the shocks to the economy (138.9). If the calculations are to be believed, the stabilization policy actually followed was much better than the totally passive policy would have been.

3.8. Taking stock

We shall end this survey of structural macroeconometric models and their application by a reflection about where we now stand in their respect, taking in particular into account the objections raised against such models and their use. We shall begin with assessments of the broad methodology by those who are directly familiar with its normal application, the author of this book included. After a brief list of the main objections and of queries about their implications, we shall discuss the trends aiming at improvement of the methodology. To this end we shall consider in succession the modelling of phenomena and the application to policy making. These various issues will be examined again from a somewhat different perspective in Section 4.6.

3.8.1. *Assessments*

We must distinguish two main purposes of the work on structural models: we may speak of the scientific purpose when referring to the testing, elucidation and development of economic theories and of the policy purpose when thinking about the models as tools for producing forecasts, specially forecasts of effects to be expected from alternative policies.

The scientific purpose clearly takes a large part in the twelve lessons that R. Bodkin, L. Klein and K. Marwah have drawn from a long experience[208]. Let us glance at extracts from their text.

1. "There does not appear to be a 'Law of Parsimony' in macroeconometric modelling... The world is complicated... simple formulas will simply not work".
2. "Although the crude accelerator, as formulated by Aftalion and J.M. Clark, does not appear to explain capital formation very well, a version of the flexible accelerator is often useful in a number of contexts".
3. "The history of macroeconometric [research] is one of slow, steady progress, rather than quantum leaps or paradigm changes – major breakthroughs are extremely rare".
4. "Financial factors have important repercussions on the real economy... conventional modelling of the real side of the world's national economies cannot encapsulate major disturbances such as world debt crises".
5. "If a disinterested party looks at all the evidence... available over the years, he... will find that the best... forecasts have been made on the basis of models in the Main Stream tradition...".
9. About the mutually reinforcing roles of macroeconometric model-building and macroeconomic theory, the authors take the example of our knowledge of the time path of the standard expenditures multiplier of the Main Stream Keynesian Theory.

Concerning the policy purpose, we may read the last line of the first paragraph in K. Wallis (1995, op. cit.): "The experience of policy analyst-advisers is that macroeconometric models provide a formal and quantified framework that is an irreplaceable adjunct to the processes of policy thought and that there is no real alternative". We may also extract the following sentence from the twelfth and last lesson of Bodkin, Klein and Marwah: "In the present state of

[208] Lessons from half a century of macroeconometric modelling, Chapter 12 in: R. Bodkin, L. Klein and K. Marwah (eds.), *A History of Macroeconometric Model-Building* (Edwar Elgar, Cheltenham, 1991).

the discipline, macroeconometric model-building is at least as much of an art as it is a set of scientific procedures"[209].

Let us note in passing that the notion according to which structural macroeconometric models are irreplaceable is specific to the policy purpose. Their importance for the scientific purpose, i.e. for the test and development of macroeconometric theories, is more debatable because there is in most cases more to learn in this respect from a less embracing framework.

We must add that unfortunately the accuracy with which are estimated the consequences of policies remains weak, particularly when the concern extends much beyond the next two years. Although looking at the details of contemplated measures is always important, doing it within very large macroeconometric models, similar to some of those built around 1970, does not much help. Ease in the resolution of models and in the introduction of occasional purpose-specific alterations is more important than whatever extra accuracy may be reasonably expected from increasing the details contained in an all-purpose model.

But learning from experience in order to improve structural models and their applications, as is suggested here, was declared inappropriate by some critics claiming that the whole approach faced strong objections and had to be rejected. There was, of course, a grain of truth in each one of these objections. But it was hasty, to say the least, to reject the approach before having the experience of alternatives which would be immune from the same objections. At this stage we shall briefly examine each one of the three main objections and draw conclusions from each.

The first was to assert that structural models were "ad hoc" and lacked microeconomic foundations. Meant as implying rejection of the models in question, this lack of foundations was proclaimed as truth and hammered over and over again, to such a point that many students were impressed and had to repeat the statement in order to ward off the fate that the objection be raised against their own productions. The objection was simply false: decent structural models have microeconomic foundations, as readers of this book have certainly realized when studying it up to this point. Admittedly, the foundations are not absolutely robust: they are not proved to perfectly apply to our economies or to rigorously imply the validity of the models at stakes. But the author of this book is still waiting to see a macroeconomic model which would have absolutely robust microeconomic foundations in such a sense, certainly not a model assuming full market clearing and applied to an economy with high or substan-

[209] This set of scientific procedures is a major concern, if not *the* major concern, of Chapter 16 presenting the twelve lessons.

tially variable unemployment, neither a model assuming a single representative agent and applied to an economy with heterogeneous households and firms.

So, if microeconomic foundations are the problem, let us seriously examine them, compare their robustness across models and see how, in each case, we might improve them. In this spirit a comparison between structural models now in use and models built by the RBC movement could be interesting.

The second main objection was proclaimed by C. Sims in his well known 1980 article, in which he proposed the vector autoregressions, the VARs, as an alternative approach intended to replace structural models for empirical macroeconomics. The objection was that the identifying restrictions of structural models were incredible (on these restrictions see Subsections 3 and 4 in Section 3.3 above). They were incredible because they were strict. As we saw, they implied, for instance, that such and such exogenous variables were not appearing in the structural equation meant to represent at the aggregate level the behaviour of some agents with respect to some of their decisions. Sims argued that practically all exogenous variables could come into play for explaining such a behaviour.

The objection looses, however, most of its strength when we consider the two following remarks: first, taking as strict a restriction that is only approximately true has little chance to much affect the quality of the representation offered by a structural equation; second, VARs too need to call on identifying restrictions and those used in practice cannot either claim more than being approximately true, as we already saw in Sections 5.5 and 5.6 of Chapter 8. So, the point made by Sims is not really a valid objection to the use of structural models but rather an inducement to compare the respective performances of VARs and structural models, a comparison which was already approached in Subsections 3.4.2. and 3.5.4 above and will be briefly examined again in Section 4.4.

The third main objection focused on the treatment of private expectations, which was said to be unsatisfactory. Indeed, as we repeatedly saw throughout our analysis of behaviours in Chapters 2 and 4, and again in Chapter 8 about the dynamics of inflation and employment, expectations were often assumed to be adaptive. We recognized, already in Section 1.6 of Chapter 8, that empirical tests of the hypothesis in question did not support it so well. In short-term analysis it may normally be the best hypothesis to make if the only feasible alternative is exact expectations, but it is clearly open to criticism.

It is particularly so in policy analysis, as was well argued in the famous Lucas' critique precisely targeted at the policy purpose of macroeconometric models[210]. Policy announcements are likely to change private expectations, particularly when they are perceived as major changes of the government's line of action. The effect of such changes is therefore poorly predicted by models assuming backward-looking adaptive expectations.

The objection often came with the suggestion to replace adaptive by rational expectations, which would so lead to alternative models, better suited for short-term analysis. But such a replacement is not a trivial operation and may materialize in different ways. Its implications for both the scientific and the policy purposes therefore needed a careful scrutiny.

We shall look, in the following Sections 4.2 and 4.3, at a number of questions raised by this alternative approach, which we briefly discussed already in Section 4.3 of Chapter 8. At this point it would be premature for us to conclude on the implications to be drawn today from the third main objection. We shall simply add here that the objection now appears to be less destructive than many macroeconomists thought twenty years ago.

3.8.2. *Modelling the phenomena*

There are permanent traits in the way macroeconomic phenomena were represented, throughout decades, by most of the then used macroeconometric models. But there were also significant changes, which should not be ignored in a complete history of these models. At this point we ought to take an overview of this history.

What will be found here is, however, brief, a fact which simply reflects the limits of the knowledge acquired by the author of the book. Indeed, giving a good historical account of the evolution undergone by the species of macroeconometric models is not easier than doing it for any living species. The reader will find historical material in the book edited in 1991 by Bodkin, Klein and Marwah. He will also be interested by what Wallis wrote, in pages 314 to 317 of his 1995 chapter, about "some recent trends in the development of the core theoretical structure of macroeconometric models".

Four traits have been maintained from the 1960s on in the macroeconometric models most used: (i) they embody aggregate demand analysis, taking into account the fact that prices and wages are imperfectly flexible (or equivalently "sticky"), (ii) they contain a dynamic representation of changes in prices and wages, in reaction to demand pressures and to changes in costs, (iii) they explicitly model changes in income distribution, as they are traced through

[210] R. Lucas, Econometric policy evaluation: a critique, in: K. Brunner and A. Meltzer (eds.), *The Phillips Curve and Labor Markets*, Carnegie-Rochester Conference Series on Public Policy, No. 1, supplement to the *Journal of Monetary Economics* (January 1976).

the income-expenditure accounting framework applied to economies in which public transfers in particular are important, (iv) they represent at the macroeconomic level fiscal and monetary policies.

Since their inception these models were less simplistic than the then prevalent, and still today resistant, popular understanding of Keynesian multiplier theory. They drew attention, for instance, to lags and foreign trade effects. It is, however, true to say that, in the early decades, quantity adjustments were seen as playing the dominant role in short-run economic movements. This was consistent with three beliefs held by most model-builders at the time: (i) the hypothesis of excess supply could be maintained in short-run macroeconomic analysis, (ii) aggregate demand had a low interest rate elasticity, demands for factors of production being viewed as fairly independent of interest and labour costs, (iii) the economy could be fine-tuned by frequent changes in the policy instruments, which would keep the economy close to a smooth and satisfactory evolution; thus, the relevant horizon for policy did not need extend beyond a rather proximate future. The main objective of the models was to correctly locate at each time the position of the trade-off between inflation and unemployment.

Already at the time these simple beliefs were not universally accepted. In particular the implementation and full effects of policy decisions were often found disturbingly long, a fact which could make fine-tuning destabilizing. M. Friedman took position against demand management activism, arguing moreover that adaptive price and wage expectations together with short-sighted policies would result in progressively higher rates of inflation without permanently improving employment (see our discussion about "the vertical long-run Phillips curve" at the end of Subsection 4 in Section 6.3 of Chapter 8).

Concerns with stagflation were indeed mounting even before they were strongly boosted by the 1973–74 shock on the terms of trade of industrial countries and the 1974–75 recession. Although the cost inflation and the depression of the world demand for goods could easily be measured and introduced as exogenous changes in the then existing macroeconometric models, which did explain the initial consequences, the subsequent evolution had to be more challenging. Policy analysts soon understood that the models were not so well suited for the study of the supply effects which would follow from the acceleration of inflation, from the drop in profitability, particularly in Western Europe, and later from the unavoidable anti-inflationary policies. Those supply effects were then perceived as involving disturbing medium-run phenomena, which had to take precedence in policy analyses over the more common short-run phenomena. At the time econometric results were also accumulating which showed that statistical fits of price and wage equations were exhibiting less robustness than had earlier been anticipated.

Thus, reasons were not lacking around 1980 for reconsidering the specifications of structural macroeconometric models, independently of the supposed fatal objections voiced against such models. Disagreeing with the thrust of these objections, teams in charge of the instruments required for policy analysis responded not by abandoning their models, but rather by amending them, seeking to remedy the various deficiencies to which recent experience had drawn attention. During this process of revision, modellers reconsidered the respective roles of microeconomic foundations and data analysis, aiming also at greater theoretical consistency.

Without going into details here, we can characterize as follows the main points of focus of this reconsideration. First, decisions of firms had to be seriously formalized, so as to give a proper and consistent representation of the joint determination of output, prices, employment and business investments, hence of labour productivity and productive capacities. This would provide the core of what can be called "the supply side". The specifications would have to bear the test of whether they were generating sensible long-run properties. Second, a closer examination of the labour market was required, both for the determination of labour supply and unemployment, and for integration of wage bargaining, even perhaps of various forms of employment status. Third, the wage-price dynamics needed a new scrutiny which would identify what could come from microeconomic knowledge and what had to come from statistical analysis of time series. Simplifying we may say that a consensus view is progressively emerging, according to which microeconomic theory has an important part to play at the specification stage, particularly because of the presence of error–correction terms embodying ideas about the long run. But the short-run time profiles have to come mainly from time-series regressions because microeconomic foundations do not give precise predictions in their respect. Fourth, the macroeconomic representation of the policy instruments and their initial impacts were reconsidered[211]. Fifth, specifications had to be clear about what was assumed, and if possible tested, about the formation of expectations. We shall consider in Sections 4.2 and 4.3 what can be now said about how to best respond to this last challenge.

3.8.3. *An instrument for policy advising*

Before closing this survey we should still reflect about the meaning and implications of assessments addressed to the policy purpose. Structural models, did

[211] See in particular F. Brayton, A. Levin, R. Tyron and J. Williams, The evolution of macro models at the Federal Reserve Board, *Carnegie-Rochester Conference on Public Policy* (1997).

we report, are said to be "an irreplaceable adjunct to the processes of policy thought". Thus, these model have to serve a function which would otherwise not be served in the processes. What is it exactly?

The use of a structural model for policy is as much an art as the application of a prescribed set of scientific procedures, did we suggest. This means that room is left for judgement. Indeed, we saw at the beginning of Section 3.4 that even the model itself frequently needs judgemental adjustments. But our function, as economists involved in policy advising, requires from us objectivity. What kind of rules, of ethical rules might we say, should we apply in this practice of judgement?

Writing about the policy purpose, we moreover found it necessary to point to the weak accuracy of structural models. Notwithstanding this weakness, we recommend to use these models, in some circumstances at least. Indeed, there are circumstances in which the predictions delivered by structural models are fairly reliable. But which are these circumstances?

Finally, we recognized the potential relevance of the Lucas' critique. On the one hand, we suggested that the point made by Lucas did not seem likely to jeopardize the normal business of policy evaluation. On the other hand, we recognized that it was challenging in abnormal cases: again under some circumstances, a government may seize the opportunity of credibly deciding an abrupt change in the whole economic policy regime. How should then the effects of the change be assessed, or alternative variants of the contemplated decision be evaluated in comparison to one another?

As we already announced at the end of Subsection 3.8.1, we shall defer to the last part of the book any attempt at answering the last question. But we shall say a few words about answers to the three preceding ones. For so doing, we shall begin with the third one, that is with the scope of the set of policy issues for which structural models are sufficiently reliable to be useful.

Facing the question makes sense. We indeed could sometimes witness cases in which, a macroeconometric model being available, its owners pretended to evaluate with it even policies for which the macroeconomic assessment had little importance. Those were structural policies, the success or failure of which depended on achievement with respect to such objectives as the development of a new industry, or increase in the productivity of technological research. Clearly the main stakes were not within the range of what builders of the macroeconometric models had envisaged.

At the other extreme a structural macroeconometric model meeting the standards of "the state of the art" should be well suited for the implementation of a contingent monetary policy rule. Such a rule will usually require something like controlling the interest rate so as to counteract the spontaneous macroeconomic trend when the diagnosis about it points to an acceleration of underlying inflation or a rise in unemployment, this at the horizon of one or two years. The policy advice then relies entirely on the short-term macroeconomic diagnosis, and this diagnosis much depends on aggregate demand analysis, for which the structural model in question is particularly well suited. Remember indeed that, notwithstanding the changes introduced in the models during the last two decades, the focus on aggregate demand at short horizons remains: the model is built on an appropriate structure for incorporating into the diagnosis direct information on private expectations and sentiments, or side analysis on such things as wealth effects, a credit crunch or a debt deflation.

So, let us repeat: when short-run changes in aggregate demand are the main concern and macroeconomic policy is not faced with disturbing medium or long-term challenges, mainstream macroeconometric models are well suited.

The potential difficulty with longer-term horizons is not so much that these models would provide a fundamentally inadequate representation of the economic system. It is rather that their quantification of medium-run effects is uncertain. The point is sufficiently important for deserving that we stop on it for a minute[212]. The difficulty of an accurate quantification has two origins. First, factors acting quickly are also generally acting directly and are well identified in the specification of the model, whereas a long lag is often the sign that the action is indirect and actually involves more than one transmission channel. In the latter case the model specification usually simplifies in assuming a more direct and unidimensional action. But estimation is then contaminated by what are really specification errors. Second, a slow action is also one that spreads over a number of periods in a way that is not a priori known. The final global impact is then less precisely estimated than if action would be felt exactly when it occurs or with a lag of just one or two periods: the sampling error in the estimation of the impact is larger, perhaps much larger.

Notice that the difficulty here stressed for the estimation of medium-run effects is inescapable. No reliable escape is provided by claiming, as is sometimes done, that the medium run is identical with the long run, to which a simpler theory would apply. However, policy advisers faced with the difficulty should be aware of it and avoid overconfidence in the results of their models.

Avoiding overconfidence is also the proper attitude with respect to the judgemental adjustments that econometricians introduce in the use of their models. Econometricians have good reasons for doing so, as we saw. These reasons are detected at times by statistical scrutiny, at times by theoretical reflection.

[212] A fuller discussion was presented in the last two pages of Section 3.1 in E. Malinvaud, Observation in macroeconomic theory building, *European Economic Review* **33** (1989) 205–223.

Statistical analysis may reveal that some actual evolutions are systematically deviating from what the model would have predicted. Reflections and investigations, such as those reported in Subsection 3.5.4, may conclude that the model is not fully appropriate for dealing with features which happen to be important in the requested diagnosis. But which exact adjustment to make is hardly ever obvious.

Since policy advisers should not be fully confident in their structural models, they should be open-minded to all kinds of evidence, in particular to that coming from alternative approaches to macroeconomic diagnosis, such as those discussed in the next part of the chapter. But they should then keep the same spirit of circumspection as they apply when considering results given by their models.

4 Developments and Alternatives

In order to complete this book we still have to consider some additional developments in the approaches to the scientific and policy purposes of macroeconomics. We shall so move still more out of the methodological domains which have been long and widely travelled, up to areas which were less extensively explored but are significant for the progress of macroeconomics. Those are areas in which research, experimentation and accumulation of evidence would be particularly welcome, also areas in which the author of this book feels more uncertain than at most other places throughout the nine chapters. Because of this last feature the treatment may be found short by the reader, who may want to gear some of his or her research to what could make it more convincing.

Five sections of this part will be respectively devoted to the search for answers to the following questions. What did we learn so far from "macroeconometric disequilibrium models"? How can we best deal with rational expectations where they are likely to play a crucial part in applied macroeconomics? What are now macroeconometric models with rational expectations? What are future prospects for the use of the VAR methodology? When and how should applied macroeconomists use calibration of the parameters included in their models? The two last sections will end this book with a more general question: considering in particular the present state of macroeconomic knowledge, how should we view academic contributions to macroeconomic policy advising?

4.1. Macroeconometric disequilibrium models[213]

Discussing structural macroeconometric models in the foregoing part we ignored attempts at the econometric implementation of the models systematically presented in Part 4 of Chapter 7, which are often called macroeconomic disequilibrium models. For reasons explained in that chapter, recognition of short-term market disequilibria extends much beyond those particular models. It is indeed a feature of most structural macroeconometric models. But when it is narrowly understood, the label "disequilibrium model" signals the presence of strong non-linearities, linked to the potential role, in comparative statics, of different combinations of market disequilibria, for instance, those distinguishing classical unemployment from Keynesian unemployment.

Particularly during the decade of the 1980s a number of macroeconometric models were built in order to find out the actual importance of the alternation between such combinations. This work had to face a number of rather thorny econometric problems, precisely linked to the inherent non-linearities of the models. Its results were instructive, as we shall see. But, in large part because of the difficulties of the econometric treatment and of the simplifications it otherwise imposed in the structure of the models, the work did not generate a new family of structural econometric models, which would be currently used by teams in charge of macroeconomic policy analysis. In other words, the econometric applications of disequilibrium models turn out to be directly useful for the scientific purpose, but not really for the policy purpose. We shall accept this verdict in considering them now.

4.1.1. *The first generation*

The first attempts, in the early 1980s, fitted models which were close to the theoretical one used in Sections 4.2 to 4.6 of Chapter 7 in this book. We may consider, as the best example here, the article by P. Artus, G. Laroque and G. Michel, looking at French quarterly data covering the years 1963 to 1978[214].

There are actually a number of substantial differences between the econometric model and the specification discussed in Chapter 7 here. First, of course, random disturbances are introduced in the behaviour equations. Moreover,

[213] For this section see G. Laroque and B. Salanié, Macroeconomic disequilibrium models, in: H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics, Macroeconomics* (B. Blackwell, Oxford, 1995).

[214] P. Artus, G. Laroque and G. Michel, Estimation of a quarterly macroeconomic model with quantity rationing, *Econometrica* (1984). See also the survey of J.-J. Laffont, Fixed-price models: a survey of recent empirical work, in: K. Arrow and S. Honkapohja (eds.), *Frontiers of Econometrics* (B. Blackwell, Oxford, 1985).

some of these equations have to cope with manifest empirical phenomena, which were neglected in our theoretical model. For instance, firm's behaviour has to be less sketchy. The presence of capital has to be recognized. Still more importantly, adjustment costs in employment have to be introduced so as to explain labour hoarding and the Okun's law; extra costs also appear when output is pushed above its optimal level given by the production function. The presence of these adjustment and extra costs implies the possibility of a fourth regime, besides Keynesian unemployment, classical unemployment and repressed inflation, a regime which the authors call "underconsumption", in which firms are rationed simultaneously on the goods and labour markets[215].

There are major differences in the representation of the labour market. The presence of frictional unemployment must be recognized. On the other hand, the supply of labour is taken as rigid and exogenous, so that it is not affected by income or wealth effects or, in the cases of repressed inflation and underconsumption, by spill-over from consumption to leisure (see what was assumed with Equations (97) and (104) of Chapter 7 about the supply of labour). Actually, starting from the labour force as given in the statistics (the sum of employment and unemployment), the share corresponding to frictional unemployment is assumed in the article to be a linear function of time, except for a random additive term; the rest is the labour supply.

Notwithstanding these differences, the basic structure of the econometric model is similar to that of the theoretical model, in that on each market, the goods market or the labour market, *exchange is equal to the minimum of supply and demand*. Since the producers on the one hand, the consumers on the other, are aggregated into single agents, they bear exactly the rationing of their supply which is imposed by a possible lack of demand, on the goods or labour market, respectively. Rationing of possible excess demand for goods falls in proportions on consumption and exports, as well as negatively on imports, which are stimulated (according to the estimates, the increase in imports is even the main effect of excess demand for goods).

In the econometric specification supplies and demands are random variables, not exactly observed. Thus, even after estimation of the parameters of the model, the values taken by those latent variables in any given period remain random, and the nature of the regime holding in that period also random: only the probability that it was Keynesian, classical, inflationary or experienced underconsumption can be estimated. Estimation of such probabilities may be taken as the most interesting outcome of the fit.

[215] The presence of this fourth regime can also be explained by voluntary accumulation of finished goods inventories or reduction of liquidity holdings.

It turns out that in most periods the estimation designates a predominant regime, the estimated probability of which exceeds 1/2. This regime is Keynesian in late 1966, in 1967 and from the last quarter of 1974 to the end of 1978 (end of the series). Classical unemployment predominates in 1963, in 1968 and early 1969, as well as in most quarters from 1971 to the middle of 1974. Repressed inflation predominates in the first semester of 1964, in the middle of 1970 and in the third quarter of 1974. The estimated probability of underconsumption exceeds one half in just two quarters (1969.4 and 1970.1).

4.1.2. *Aggregation of micromarkets*

In applied disequilibrium macroanalysis the argument for looking at aggregation issues is compelling. Indeed, it seems hardly satisfactory to rely on a theoretical model organized around "the minimum condition", which states that the minimum of supply and demand is traded in the aggregate. That this condition lacked realism was the main message of Section 4.8 of Chapter 7. It was there shown that the multiplicity of markets experiencing different types of disequilibria had the effect to smooth the relationship linking aggregate output to the main determinants of supplies and demands. But that section did not go far into the discussion of models which would correctly describe the underlying microeconomic phenomena and their aggregation. This had to be studied as a preliminary to the specification of good econometric models.

In order to exhibit the main steps leading to aggregate specifications we may borrow from G. Laroque and B. Salanié (1995, op. cit.) the following short presentation concerning the case of a single macromarket, "the market for goods" say (references to the supporting theoretical literature is to be found in that article).

Let there be N elementary commodities, each traded on a specific micromarket ($n = 1, 2, \ldots, N$). For each of these commodities, there are suppliers and demanders, whose behaviour at predetermined prices is summarized through supply and demand functions, which depend on a number of variables: (i) the quantities traded on the other markets, of which the $(N-1)$ vector is denoted as q^{-n}, (ii) a vector of exogenous variables x, (iii) a vector of parameters θ and (iv) a vector of random shocks ε (at this stage, the vectors x, θ and ε have to be seen as very large, since we do not exhibit that some of the exogenous variables, parameters and shocks have to be specific to particular micromarkets). We let:

$$d_n = D_n\big(q^{-n}, x, \theta, \varepsilon\big), \tag{238}$$

$$s_n = S_n\big(q^{-n}, x, \theta, \varepsilon\big), \tag{239}$$

where d and D (respectively s and S) stand for demand (respectively supply). The minimum rule is assumed to hold on each micromarket, so that the traded quantity on market n satisfies:

$$q_n = \mathrm{Min}\{d_n, s_n\}. \tag{240}$$

At this stage the model allows for spill-over effects, since the traded quantities on markets other than n (the components of vector q^{-n}) appear as potential determinants in the demand and supply functions on market n. What becomes of these spill-over effects in the aggregate? The few models that have studied the issue yield aggregate specifications that, first, do not differ very much from the models without spill-over effects, and second, have not led to empirical applications. As a consequence, it has become customary to concentrate on a model without such effects, where the aggregate quantity is simply the sum of the quantities traded on the elementary markets, and where specific assumptions are made about the demand and supply functions, as well as about the shocks. Let us look at this much more special model.

Two kinds of shocks are assumed to occur on the micromarket n: some idiosyncratic shocks represented by ε_{nd} and ε_{ns}, affecting respectively demand and supply, and some common aggregate shocks ε_d^a and ε_s^a. Except for the effect of idiosyncratic shocks, supply and demand are assumed to be the same on all markets, so that (238)–(239) are specialized as:

$$N \cdot D_n\big(q^{-n}, x, \theta, \varepsilon\big) = D\big(x, \theta, \varepsilon_d^a, \varepsilon_{nd}\big), \tag{241}$$

$$N \cdot S_n\big(q^{-n}, x, \theta, \varepsilon\big) = S\big(x, \theta, \varepsilon_s^a, \varepsilon_{ns}\big). \tag{242}$$

The minimum rule (240) then gives the aggregate quantity traded:

$$Q = \frac{1}{N}\sum_{n=1}^{N} \mathrm{Min}\Big\{D\big(x, \theta, \varepsilon_d^a, \varepsilon_{nd}\big), S\big(x, \theta, \varepsilon_s^a, \varepsilon_{ns}\big)\Big\}. \tag{243}$$

Moreover, in empirical applications, the two kinds of shocks are assumed to appear additively in the demand and supply functions, the idiosyncratic shocks are assumed to be independent from the aggregate shocks and, conditionally on the latter and on the exogenous variables, the idiosyncratic shocks are also assumed to be independently and identically distributed across markets. Since the number of micromarkets is seen as large, we may then call upon the law of large numbers and translate the right-hand member of (243) into a conditional mathematical expectation taken over the probability distribution of idiosyncratic shocks:

$$Q = E\Big[\mathrm{Min}\Big\{D\big(x, \theta, \varepsilon_d^a, \varepsilon_{nd}\big), S(x, \theta, \varepsilon_s^a, \varepsilon_{ns})\Big\}/x, \varepsilon_d^a, \varepsilon_s^a\Big]. \tag{244}$$

This is a function of x, θ, the aggregate shocks and the parameters of the distribution of idiosyncratic shocks (to be added as components of θ). Finally, the proportion of markets in excess demand is equal to the conditional probability:

$$p_{D>S} = \text{Prob}\Big[\Big\{D\big(x, \theta, \varepsilon_d^a, \varepsilon_{nd}\big) > S\big(x, \theta, \varepsilon_s^a, \varepsilon_{ns}\big)\Big\}/x, \varepsilon_d^a, \varepsilon_s^a\Big]. \tag{245}$$

Equation (244), defined and interpreted as above, gives the macromodel for the case of a single macromarket. Going through the same successive steps econometricians can find ways of defining macromodels corresponding to cases of two or several marcromarkets. Equations similar to (244) then define the macroeconomic endogenous variables as functions of exogenous variables (x), parameters (θ) and aggregate shocks. Econometric estimation may conceivably give values of the parameters to be inferred from observed time series of the aggregate endogenous and exogenous variables, as well as from an hypothesis about the nature of the probability distribution of aggregate shocks.

The reader will not be surprised to learn that finding appropriate estimation methods in this framework was challenging econometric theory. Actually, the relevant functions connecting the observations to the parameters typically do not even have analytical expressions. To get around the difficulties, the econometrician can, however, use Monte Carlo simulations and focus attention on the functions giving how the first order, and possibly second order, moments of the probability distribution of the aggregate endogenous variables depend on the exogenous variables and the parameters[216].

Following this approach leads moreover the econometrician to better understand how the disequilibrium model with micromarkets observationally differs from the corresponding model specified directly at the macroeconomic level without micromarkets, i.e. the model assumed in the first generation of disequilibrium econometric investigations (see the preceding subsection). Indeed, it is found that the mathematical expectations of the aggregate traded quantities only depends on the parameters of the distribution of sums such as $\varepsilon_d^a + \varepsilon_{nd}$ and $\varepsilon_s^a + \varepsilon_{ns}$: in other words, assuming $\varepsilon_{nd} = \varepsilon_{ns} = 0$ as was done in Subsection 1 makes no difference for the results to be reached by any estimation which will consider just those relations between observations and parameters which are implied by consideration of the first order moments.

G. Laroque and B. Salanié (1989, op. cit.) wanted to know which difference it made in practice to recognize the existence of micromarkets. They therefore looked at the results that could be obtained from fits of the models to a

[216] On this point and what follows, see G. Laroque and B. Salanié, Estimation of multi-market fix-price models: an application of pseudo maximum likelihood methods, *Econometrica* (July 1989).

data base of French aggregate quarterly series. For parameters other than those concerning the distribution of disturbances, the results were quite close across estimation procedures, as well as between the purely aggregate specification and that admitting micromarkets. The authors also noted that introducing spatial disturbances (between micromarkets) and estimating them through their impact on the second-order moments of endogenous variables did not seem to induce any significant reduction in the estimated size of the time disturbances. Estimation of the variance of the spatial disturbances on the labour market gave a small but very imprecise result. Spatial disturbances on the goods market were estimated to have a standard error equal to 1.25 times that of time disturbances (this number 1.25 being itself subject to a sizable standard error: 0.45).

Overall this investigation suggests that the estimations obtained in the first generation of econometric works about the disequilibrium models are robust with respect to the explicit introduction of micromarkets. Such a conclusion could not be taken for granted in advance.

4.1.3. *Using extra information about market tensions*

Up to this point the estimation methodology discussed uses only data on traded quantities and on the classical determinants of supplies and demands. The main difficulty then stems from the latent variable problem. Getting supply and demand estimates would be easy if the prevailing regime was known at each date: one would just have to be careful and beware of the fact that the distribution of the residuals of supply and demand equations, conditional on the observed regime, differs from the unconditional one.

We may think that the main advantage of the micromarkets specification is to open the possibility of using a wider range of empirical evidence, including data such as those provided by disequilibrium indicators. We would thus alleviate the latent variable problem. Intuition suggests that the information content of the indicators in question is relevant in estimation, particularly for taking advantage of equations such as (245), which was derived from the theoretical model but not used in the methodology sketched above.

The main available indicators of market tensions or slacks come from the statistical processing of what market participants say: the households for the rate of unemployment, the industrial firms for the rate of utilization of the productive capacity of equipments, and for the proportions of firms that are constrained by either lack of sufficient demand for their output, or lack of sufficient equipment, or still difficulties in recruiting adequate personnel. The practice with the collection of such data is now old enough to have been standardized, at least within countries, and to have delivered time series which are reliable in themselves, even though their interpretation in terms of economic concepts may be still disputable in some cases (Laroque and Salanié, 1995, op. cit., have a short discussion on this last point and suggest that the presence of "errors on variables" should be recognized when, for instance, the proportion of manufacturing firms, weighted by their sales, who report recruitment difficulties is taken as a measure of the proportion of labour micromarkets experiencing excess demand).

For econometricians, aware of the many problems they are facing, the information given by such indicators of market tensions as experienced by economic agents cannot be taken as negligible. Indeed, it was not fully neglected in structural macroeconometric models, but it was then introduced in one or two equations without a systematic discussion of its insertion into the model. The relevant question for us at this stage is indeed to know how to best use the information in question. For the time being there is just one model within which it has been systematically introduced for the characterization of disequilibria in the changing temporary macroeconomic equilibrium of national economies. This model is definitely restrictive but, according to the author of this book, roughly appropriate to the characterization at stake. Let us briefly see what the problem is and how it is solved in that particular model[217].

The problem is here again the recognition of spill-over effects in micromarkets, in particular from labour markets to goods markets or the reverse, this being modelled in such a way as to lead to a fairly simple specification at the aggregate level. Those who looked into the problem found it intractable in general specifications. The only available solution thus far consists, on the one hand, in strictly coupling the labour and goods micromarkets and, on the other hand, in choosing strict and convenient assumptions about the occurrence and distribution of idiosyncratic shocks.

Let us assume that the micromarket n has two corners: the goods corner in which demand d_n and supply s_n meet, the labour corner in which demand ld_n and supply ls_n similarly meet. Let us moreover assume that firms operating on this market use a putty-clay technology so that, in the short-run temporary equilibrium, output q_n and employment l_n have to be proportional: $l_n = A_n q_n$, with a given number A_n; moreover output cannot exceed the standing productive capacity of equipments qc_n.

Taking into account the minimum conditions on the two corners of the market we see that three well characterized fixed-price equilibria could hold. Firms

217 At the end of their Part 2, Laroque and Salanié (1995, op. cit.) hint at an approach that would allow more flexibility in the specification and might be explored.

might be rationed on the labour corner (repressed inflation):

$$l_n = ls_n, \qquad q_n = ls_n/A_n = q_n^r \tag{246}$$

or they might be rationed by lack of demand on the goods corner (Keynesian equilibrium):

$$q_n = d_n = qk_n, \qquad l_n = A_n qk_n = lk_n \tag{247}$$

or still they might hit the capacity limit (classical equilibrium):

$$q_n = qc_n, \qquad l_n = A_n qc_n = lc_n. \tag{248}$$

In any case the two following minimum conditions simultaneously hold:

$$q_n = \text{Min}\{qr_n, qk_n, qc_n\}, \qquad l_n = \text{Min}\{ls_n, lk_n, lc_n\}. \tag{249}$$

This equilibrium classification being given, we understand that we must say how ls_n, d_n, qc_n and A_n depend on variables, parameters and shocks. We also anticipate that, in the route to building a macroeconomic model, we shall find steps similar to those which led us from (238)–(239) to (244)–(245). Without stretching too much our imagination, we may guess that functions will be found which will explain, at the aggregate level, not only the variations in labour supply LS, aggregate demand YD, capacity output YC, the labour-output coefficient A, but also the proportions π_r, π_k and π_c of micromarkets respectively in the three kinds of equilibria defined above, and other relevant variables as well. We may also expect that the growth rate of aggregate output $\dot{y}$ will then be (approximately) equal to

$$\dot{y} = \pi_r g(LS/A) + \pi_k g(YD) + \pi_c g(YC), \tag{250}$$

where $g(x)$ is, by definition, the growth rate of x. We shall not look at each step of this derivation[218], but rather focus on what is assumed about idiosyncratic shocks in order to make the macroeconomic model tractable and transparent.

The main point is the form of equations similar to (241)–(242). Omitting the aggregate arguments of the functions, we may write these equations as:

$$N \cdot ls_n = \varepsilon_{lsn} LS, \qquad A_n = \varepsilon_{an} A, \tag{251}$$

$$N \cdot d_n = \varepsilon_{dn} YD, \qquad N \cdot qc_n = \varepsilon_{qcn} YC. \tag{252}$$

The idiosyncratic shocks ε_{lsn}, ε_{an}, ε_{dn} and ε_{qcn} are assumed to be independently and identically distributed across micromarkets, and to be independently distributed of aggregate shocks and exogenous variables. They now

[218] See in particular J.-P. Lambert, *Disequilibrium Macroeconomic Models: Theory and Estimation of Rationing Models using Business Survey Data* (Cambridge University Press, 1988); J. Drèze and C. Bean, Europe's unemployment problem: introduction and synthesis, in: J. Drèze et al. (eds.), *Europe's Unemployment Problem* (MIT Press, 1991).

occur multiplicatively with mathematical expectations equal to 1. Their probability distribution is moreover strictly assumed to be made of four independent Weibull distributions, the variance of ε_{an} being equal to that of ε_{lsn} and the variance of ε_{qcn} equal to that of ε_{dn}.

J.-P. Lambert (op. cit.) proved that these assumptions lead to pleasant forms for aggregate output and aggregate employment. For instance, the mathematical expectation L of l_n taken over the probability distribution of idiosyncratic shocks gives:

$$L^{-\rho_1} = (LS)^{-\rho_1} + A^{-\rho_1}\big[(YD)^{-\rho_2} + (YC)^{-\rho_2}\big]^{\rho_1/\rho_2}. \qquad (253)$$

This is an appealing expression[219], a form of generalized CES function (CES meaning "constant elasticity of substitution"). The positive parameter ρ_1 is a decreasing function of the variance of ε_{an} or ε_{lsn}, varying from ∞ to 0 as the variance goes from 0 to ∞ (note that $\rho_1 = \infty$ would make L equal to the minimum of LS and A multiplied by the square bracket to the power $-1/\rho_2$, in other words the aggregate demand for labour). Similarly ρ_2 is a decreasing function of the variance of ε_{dn} or ε_{qcn}. The number $1 - 2^{-1/\rho_1}$ can be taken as a measure of mismatch in the distribution across micromarkets of the labour supplies and demands; indeed, it would be the unemployment rate according to (253) if the aggregate labour supply and demand were equal. A similar measure of the "structural underutilization rate of capacity at equilibrium" can be computed from ρ_2.

Within a model of the type so described, business survey data were used, in conjunction with macroeconomic data, not only for a few national studies but also for a systematic econometric analysis of employment trends in most countries of Western Europe and in the US. J. Drèze and C. Bean (op. cit.) publish a synthesis of the results reached in that international analysis[220]. In particular they provide yearly estimates, from 1965 to 1986, of the proportions π_k, π_c and π_r of micromarkets where domestic output is determined by demand, by capacity and by availability of labour respectively. Little trend appears in the US, π_c moving procyclically and π_k countercyclically. But in Europe π_k grew markedly from 1975 and especially 1981 onwards. The levels of these proportions are, however, quite different from one country to another, a fact which suggests a poor international comparability of business survey data (in 1980, π_k is found equal to 0.86 in Britain, 0.39 in Germany and 0.29 in the US; for

[219] However, commenting on the assumptions leading to it, Laroque and Salanié (1995, op. cit.) write "The price to pay in order to keep a simple closed form seems to be rather high, in terms of *ad hoc* assumptions".

[220] The conclusions about the persistence of European unemployment were presented in Section 2.9 of this chapter.

the same year and the same three countries π_c is respectively equal to 0.05, to 0.38 and to 0.53). The decomposition of the output growth rate according to Equation (250) shows, for most European countries, negligible contributions of capacity and labour supply growth in the eighties, in contrast to the US where these contributions retain their importance.

In the fit of Equation (253) the exponents ρ_1 and ρ_2 were assumed to change according to linear trends of their inverses (and to have equal values for all countries except France). In conformity with the outward shift of the Beveridge curve, which was observed elsewhere by many macroeconomists[221], the value of $1/\rho_1$ is found to have increased significantly in all countries, leading for instance in Germany to a "structural unemployment rate at equilibrium" equal to 1.9 per cent in 1960 and to 4.6 per cent in 1986. The French result and other evidence for some countries lead the authors to argue that no substantial trend exists for the "structural underutilization rate of capacity at equilibrium".

4.1.4. *Dynamic disequilibrium models*

Business survey data are not unique as an additional source of information which might be available for improving upon the results reached by the first generation of works on macroeconometric disequilibrium models: a particularly rapid increase in prices (alternatively wages) may signal presence of excess demand in the goods market (alternatively the labour market). Without even speaking of micromarkets, econometricians may try to incorporate the relation between price changes and excess demands into their disequilibrium models. They may do so not only because price adjustment equations are interesting *per se*, but also because the presence of these equations should provide a kind of automatic control that is lacking in static disequilibrium models. Indeed, it may be the case in such static models that the fit implies excess demand at a time when price increases were small with no other sensible reason than excess supply; perhaps this would be explained by the fact that estimates of some of the demand or supply parameters went astray, for instance supply parameters if few periods of excess demand were observed in the sample; there would then be a good chance that addition of price and wage-change equations in the disequilibrium model would have prevented such misestimation of the supply parameters in question.

Disequilibrium macroeconometrics may have other reasons for studying dynamic specifications. For instance, we should like to know which part, if any,

[221] Some macroeconomists, however, questioned the idea that this shift was caused by an increase in the structural mismatch and wondered whether it was not due to a decline in the intensity of job search by unemployed people.

is played in persistence of unemployment by the possible sluggishness of the productive capacity of equipments. Under which conditions would adaptation of this capacity be too slow in its reaction to bursts of aggregate demand and would then prevent the full and quick impact on output? Has the level of profitability anything to do with such conditions?

The study of dynamic disequilibrium models is still little developed. It has indeed to face a number of difficulties, concerning both the choice of adequate specifications and their econometric treatment. These difficulties are presented in Part 3 of the 1995 article written by G. Laroque and B. Salanié (op. cit.). They are related in particular to the role of lagged latent variables, such as lagged demands and supplies, in models which have otherwise the non-linear structure imposed by the minimum condition(s) on some endogenous variables such as output. We shall not enter here into the field which is open for the study in question, except for a brief presentation of a result that confirms the ideas sketched in the first paragraph of this subsection.

B. Salanié[222] fitted quarterly French data for the years 1964 to 1984 to a model consisting of two blocks: a "real block" that takes the current wage rate and price level as predetermined and determines employment, output and the main components of aggregate demand, a "nominal block" that yields the rates of growth of price and wage levels between quarters t and $t+1$, as a function of excess demands on both markets in quarter t. Since the model has a recursive structure, it is possible to consistently estimate the real block, either in isolation (as would have been done in the first generation of works discussed in this section) or within the full model made of the two blocks. The values estimated in the two approaches for the parameters of the real block do not much differ, but their precision is found to be higher with the second approach: nominal equations seem to bring in useful information about the variables explained in the real block. We note finally that the estimated excess supply of labour is very highly correlated with the official unemployment figure and that the estimated wage equation is consistent with what had been found for the French Phillips curve by various authors who used a different methodology.

4.2. The rational-expectations approach to applied macroeconomics

Any modern macroeconomist is likely to have views about the proper way of using in his or her discipline the hypothesis of model-consistent expectations.

[222] B. Salanié, Wage and price adjustment in a multimarket disequilibrium model, *Journal of Applied Econometrics* **6** (1991) 1–15.

At various earlier places in this book material was presented and comments were made in order to help the reader to form his or her judgement in this respect. But more must now be added. A good reason for doing so may come from the feeling that there is today less fundamental disagreement in the profession about the issue than was the case twenty years ago. Another reason comes from the experience gained by those who introduced the hypothesis in structural macroeconometric models. Although the implications of this experience will probably better be seen in a decade than now, we must try and assess them as well as we can. We must also provide the reader with elements of the technical knowledge necessary for working on macroeconometric models with rational expectations.

With these purposes in mind we shall, first, survey the challenge posed to the discipline by the interaction between changes in expectations and changes in the macroeconomic state. We shall, second, discuss the Lucas' critique. We shall, third, briefly look at the formal theory of linear rational-expectations models. We shall, fourth, examine how to proceed to policy analysis in this linear context. We shall, finally, come back once again to the analysis of a monetary policy rule, but now in an economic environment where forward rational expectations would dominate.

4.2.1. *The challenge*

All economists "would agree that the response of the macroeconomy to disturbances will depend on the beliefs, perceptions and expectations of participants". But R. Solow[223], the author of this undisputable sentence, adds: "I would not know what set of statements about these things deserves to be included in the core of usable macroeconomics. I feel acutely uncomfortable with this fudge factor that is capable of having drastic effects but is so conjectural that it can be used to explain just about anything". Probably a bit too extreme, these latter two sentences well pose, however, the importance of the challenge to macroeconomics that concerns the knowledge of how are formed beliefs, perceptions and expectations.

In order to respond macroeconomists have, in the first place, to carefully look at what is currently published about the state of "these things" in the economy they are considering. They also have to request that more information on it be collected, because what is available hardly ever do justice to the crucial role of the "fudge factor". But observing is not enough, we must also

[223] R. Solow, in the symposium, Is there a core of practical macroeconomics that we should all believe?, *American Economic Review* (May 1997).

explain in order to be able to make comparative assessments of policies: explain the formation of perceptions and expectations, explain how much they change after the announcement of such or such change in policy, explain their roles in behaviour. The branch of econometrics dealing with such phenomena is regrettably little developed.

We know, however, that, in a number of cases, data disproved the hypothesis that expectations would have been adaptive. It is unfortunate for the analyst, because models with adaptive expectations deal with successive periods sequentially, a feature which makes the study of optimal dynamic control easy in principle. If the endogeneity of expectations matters and if the hypothesis of adaptive expectations would, more or less frequently, lead to specification errors, *what should macroeconomists do*?

The rational-expectations movement claimed to provide economists with theoretically grounded models which would explain in particular the formation of expectations. In the 1970s many proponents had no doubt that correct explanations would so be obtained. After more than two decades we can see that this extreme claim was not found to be empirically supported. We moreover see, in particular when reading the symposium published in May 1997 by the *American Economic Review*, that opinions still very much vary about what the macroeconomists should do.

J. Taylor writes: “People’s expectations are highly responsive to policy and thus expectations matter for assessing the impact of... policy. The most feasible empirical way to model this response is the rational expectations approach, though modifications to take account of differing degrees of credibility are necessary[224]. By introducing rational expectations into fully estimated econometric models and then simulating the models for different policies, the response of expectations to changes in policy can be reasonably approximated”.

For his part R. Solow writes: “The main, perhaps the only, merit of the rational-expectations hypothesis in the macro field appears to be its definiteness... I can see a role for rational expectations in modeling of long-run equilibrium. In the short-run part of macroeconomics, the rational-expectations hypothesis seems to have little to recommend it. In that context, I suspect that expectations are best handled ad hoc, that is, in a commonsense way. This takes self-discipline. If the danger with the rational-expectations hypothesis is that it is too often definitely wrong, the danger with my suggestion is that it may be vacuously right”.

[224] J. Taylor had earlier written: “There is no agreed-upon way to adjust these new econometric models to deal with credibility”. He had then presented his somewhat ad hoc adjustments in a study of a multiperiod plan to reduce the US budget deficit. See, The use of the new macroeconometrics for policy formulation, *American Economic Review* (May 1993).

A. Blinder is quite short on the issue of "modeling expectations" and does not really conclude. But he points to the meager empirical success of what was called the rational-expectations revolution. "Where expectations can be measured directly, they do not appear to be "rational", as economists use that term [see here, Section 1.6 of Chapter 8]. And at least some empirical relationships, including the term structure, seem to work better with adaptive than with rational expectations[225]".

Even though we cannot but feel uncomfortable, we shall try and speculate about which common wisdom might emerge among economists using structural macroeconometric models for policy analysis. The common wisdom would be aware of doubts about the realism of the hypothesis, but also of the often poor empirical performance of the only available competitor: the hypothesis of adaptive expectations. The common wisdom would also be aware that rational expectations may entail heuristically strange consequences (see Section 4.3 of Chapter 8). It would moreover be aware of the logical and practical difficulties met by those who earlier worked on models giving an important, perhaps even ubiquitous, role to rational expectations (see Subsection 4.3.3 below). It would finally have a correct evaluation of the bearing of the Lucas' critique (see below).

The future common wisdom may turn out to be something like the following: before replacing adaptive by rational expectations about a variable, reflect on whether the substitution is likely to mean a gain in realism; before working out the complete macroeconometric model, try and make simulations on a small-size maquette, which will give you a sense of how solutions of the full model are likely to behave; be aware of the experience showing that the risk of seeing your analyses marred with unstable solutions seems to increase with the number of variables for which rational expectations are assumed; in policy analyses pay great attention to the likely degrees of credibility of alternative policy packages; reflect on how to represent in your model policies which economic agents are not likely to find fully credible; reflect also on how to later make adjustments in the results delivered by the model about the effects of such imperfectly credible policies.

4.2.2. *The Lucas' critique*

It is now the proper time to discuss in general terms the implications of the Lucas' critique. The critique was one of the points made by the rational-expectations revolution, but we may, to begin with, see it as being even broader

[225] G. Chow, Rational versus adaptive expectations in present value models, *Review of Economics and Statistics* (August 1989).

and as challenging the use of models in policy analysis, even independently of what the models may assume about the formation of expectations[226].

Indeed, a basic hypothesis is always made when a model is applied in order to predict the effects of a contemplated policy, namely that the model is applicable to the case. More precisely the model is ready to be used, in particular with values given to its parameters (for definiteness we may speak of "the model in its quantified form", as distinct from "the general form" in which values are not yet given to parameters); the policy is defined as a rule for the determination of instruments appearing as exogenous variables in the model; forecasts are made on other exogenous variables in the model (in Subsection 3.6.1 we proposed to call them "variables on surrounding conditions"); solution of the model gives the prediction of endogenous variables. This was displayed for linear structural models in Section 3.6, but is much more general. With appropriate definition of the parameters, we may even use this presentation for dealing with cases in which the general form of the model covers several regimes, values of the parameters and the exogenous variables telling which regime applies. Policy analysis compares the effects predicted as resulting from alternative policies or from alternative modalities of the same policy, the values of parameters and variables on surrounding conditions being kept unchanged. Thus, the basic hypothesis about the applicability of the model may be said to be the *invariance of parameters*.

Lucid model builders and users have the constant worry of knowing whether this invariance is likely to hold. It is indeed a worry because only perfect knowledge could provide sure answers, but nobody has perfect knowledge. The worry is present from the beginning when the general form of the model is defined, because ideally it should be such as to guarantee parameter invariance at later stages. It is also present when the question is raised of knowing whether and how estimated values of parameters ought to be revised for a new set of applications. An so on[227].

Writing θ as the notation for characterizing the set of parameters that are claimed to be invariant in an application and aiming to later test the Lucas' critique, Favero and Hendry distinguish three levels of a critique that can be addressed to the claim of invariance:

[226] This broader perspective is wisely taken at the beginning of C. Favero and D. Hendry, Testing the Lucas critique: a review, *Econometric Reviews* **3** (1992) 265–306.

[227] Parameter invariance is not the only source of worry of model builders and users. A similar and related problem was often discussed about structural macroeconometric models, namely whether their various structural equations were not given too much autonomy when they were gathered into a model of the economy and maintained unchanged in applications, unless there was for any of those equations a quite precise reason for revising it. For the history of the concept see J. Aldrich, Autonomy, *Oxford Economic Papers* **41** (1989) 15–34.

(a) θ will vary directly with changes in the environment;
(b) θ will vary with changes in economic policy control rules;
(c) θ will vary with changes in the environment which alter expectations.

In this list the expression "the environment" does not refer to what is captured by variables on surrounding conditions, but to all the rest, including the context within which policy rules are chosen or revised. So, critique at level (a) may be so general as to be distressingly unhelpful. Favero and Hendry write that such a critique "is intended to entail the claim that there may exist sufficiently large changes such that empirical evidence embodied in econometric evidence ceases to be relevant. This level reflects the fact that total invariance of econometric models to all possible changes in the environment cannot be achieved in a social science". Clearly, such a critique would give a too easy argument to anyone who decided, for whatever other reason, to disparage a serious examination of an economic policy. It happened to this author to read invocations to the Lucas' critique which did not mean more.

Critique at level (b) may also be quite general, hence again unhelpful, or it may be specific. The main point of listing level (b) is to lead the reader to realize that a specific criticism to parameter invariance does not need to involve expectations. It may concern other aspects of behaviour, or features of market adjustments which were assumed in the specification of the model or which held in all the evidence used for the quantification of parameters, but would be affected by a contemplated policy rule. Such a specific criticism may focus, for instance, on the assumed "autonomy" of an equation, which ought to be revised before application of the model.

R. Lucas in his famous 1976 article went as far as claiming that "any change in policy will systematically alter the structure of econometric models" (p. 41). As such, the sentence can only be wrong, because it is disproved by any single example concerning a change of policy that does not alter the structure of an econometric model meant to be relevant for the analysis of this policy. Clearly Lucas was not referring to any change in the values of the policy instruments, but rather to any change in the policy rule, that is, excluding routine changes in the instruments, changes which are not often at stake in policy evaluations (this is clear from the paragraph following the quoted sentence in Lucas' article). Even so, the statement is too embracing and systematic to be true.

The explanation, given by Lucas in support of the statement is, however, interesting because of the role it gives to expectations. The critique of parameter invariance concerns level (c) of Favero and Hendry. Slightly rephrasing the explanation, we can state it as made of four assertions:

(1) the structure of an econometric model contains a representation of the decision rules followed by economic agents and ought to systematically vary when these rules change;
(2) the decision rule followed by economic agents systematically vary with the way in which agents' expectations about the variables relevant to them are formed;
(3) the formation of private expectations systematically varies with the structure of the process generating the time series of these variables;
(4) any change in policy will alter that process.

Actually what we call here sentences 2 and 3 are more precise in Lucas' article, where they assert full rationality of private agents: the decision rules in 2 are optimal; 3 refers to rational expectations.

Here we should reflect on the meaning and implications of the argument, rather than on its exact phrasing, which is secondary and disputable (again because it is too embracing and systematic). We shall have more to say about the succession of the four sentences at the end of Subsection 4.2.4. But, at the present level of generality, we should be ready to recognize the presence of potential difficulties, which might have seriously biased the results of some past policy evaluations. We should then like to see good examples in which such biases occurred, precisely for the reasons exhibited in Lucas' article. We should further wonder how a better job could have then been done by policy analysts, assisted by econometricians.

Indeed, nothing in the spirit of the macroeconometric movement prevents in principle to define models with a sufficiently encompassing structure for it to accommodate a relevant diversity of private decision rules, particularly if we know how to relate those to the dynamic properties of the process of macroeconomic variables. Nothing in principle prevents us from explicitly introducing expectations, either as observed or as latent variables. For progressing in our macroeconomic knowledge and for being in a better position to advise policy makers our problems are more practical. They are to learn from our past failures, or even may be successes. They are to experiment along directions which were not enough explored.

The tests presented by Favero and Hendry (op. cit.) are interesting as showing how to learn about possible invariance failures, at least in small-size econometric models. The authors also present an application for a test at level (c), which turns out to be unable to reject parameter invariance in an econometric model of the US transaction demand for money: in the model no forward-expectation variable appears, whereas such variables might be thought to have played a significant role, considering the changes that occurred in the money supply process during the period (in Subsection 4.2.4 we shall be able to understand the nature of the test). The authors conclude: "The onus seems to have shifted from [econometric] models having to prove their innocence from the Lucas critique to critics having to empirically establish their guilt".

But, as was well shown by K. Wallis, it is also possible to introduce rational forward-looking expectations in the specification of structural macroeconometric model and to so device a practical response to the Lucas' critique. This is precisely what we are going to study in the two following subsections.

4.2.3. *Linear rational-expectations models*

For simplicity we often focused on the linear case in discussing various aspects of structural macroeconometric models. We shall do so again now, while considering the basic new elements coming from the introduction of rational expectations (discussing the experience of the last two decades in the next section, we shall of course not stick to linear models).

Within the rather narrow linear category we shall even refrain from looking at general formalizations, which quickly become heavy. Our approach will lead us to select dynamically simple models and even to abstain at the beginning from exhibiting the instrument variables, which will have to occur in any policy evaluation. Our presentation will be directly inspired by the first article in which K. Wallis showed how to deal with rational expectations in structural models[228].

1. Following Wallis we begin with the case in which *past expectations about current values* affect the current equilibrium, but present expectations about future values do not. This is not the most interesting case, but one which is simple and already exhibits important features of rational expectations models.

The model contains a vector x_t of m observable exogenous variables a vector y_t of n observable endogenous variables and a vector η_t of n unobservable disturbances. The new feature is the presence in structural equations of the expected or anticipated values of some endogenous variables. For simplicity at this first stage they are elements of the vector y_t^e expected at time $t-1$ to hold at time t. More precisely writing J_{t-1} for the information available at time $t-1$ and E for the mathematical-expectation operator we pose:

$$y_t^e = E(y_t/J_{t-1}). \tag{254}$$

The notation is identical to that introduced in Section 1.5 of Chapter 8. But, in the framework to be now discussed, the equation has a strong theoretical

[228] K. Wallis, Econometric implications of the rational expectations hypothesis, *Econometrica* (January 1980).

content: it assumes a particular theory of expectations formation, indeed at least a first part of the "rational-expectations hypothesis" as usually formalized.

Except for the presence of y_t^e the model is static:

$$By_t = Dy_t^e + Ax_t + \eta_t \tag{255}$$

with the three matrices of coefficients B, D, A (possibly containing many zero elements), the matrices B and $B - D$ being non-singular. The unobservable disturbance vector η_t is assumed to have zero expected value. Its process is usually assumed to be white noise.

We might consider (255) alone as providing "the model". For instance, we might observe the relevant part of y_t^e, the part made of the elements j for which the corresponding column of D is not made of zeros only; and we might moreover take the expectations to be exogenous. Or we might take the process $\{y_t^e\}$ to be unobservable and purely random, in which case D_t^e could be merged with η_t and we would be back to the usual formalization of structural models (with the simplifications accepted here). But our model assumes that y_t^e is non-observable and ruled by (254): these two assertions are part of the model.

Can we derive from (254) and (255) an equation involving only observable variables and disturbances? A natural idea is to take mathematical expectations, conditional on J_{t-1}, on both sides of (255) and to derive:

$$(B - D)y_t^e = AE(x_t/J_{t-1}) = Ax_t^e \tag{256}$$

leading to:

$$By_t = Ax_t + D(B - D)^{-1}Ax_t^e + \eta_t. \tag{257}$$

The operation is not innocuous: deriving (256) is indeed using the second part of the rational-expectations hypothesis, namely that the expectations are model-consistent: economic agents know the model (in its quantified form, it is part of J_{t-1}). Again we could stop at (257) if the expectations about exogenous variables were observed. But in normal applications they are not. We therefore need a theory explaining these expectations.

The natural reaction of rational-expectations economists is then to look at *the* model which describes the determination of the values of exogenous variables, hence to add a "complement" to what we might have thought to be the structural model, namely (256)–(257): rational expectations of the values of exogenous variables also have to be model-consistent. We must grant that there is nothing fundamentally wrong in this reaction: if expectations about exogenous variables matter and if they are not observed, they have to be somehow explained. The point of the remark is rather to show, first, that the scope of model-consistency has to be wide, second, that the specification assumed in order to "complete the model" may be as important, even perhaps more important, than the specification assumed about the determination of endogenous variables, here (255).

In conformity with what was usually assumed by rational-expectations economists, Wallis postulates that the vector of exogenous variables is ruled by an ARMA model. It will not be much more restrictive, but be simpler, to postulate here the following autoregressive process:

$$x_t = \Phi(L)x_{t-1} + \xi_t, \tag{258}$$

where the unobservable disturbance vector ξ_t follows a white noise process independent of the one ruling η_t, and where $\Phi(L)$ is a finite polynomial of the lag operator L (for simplicity we ignore presence of a constant term). Accepting the assumptions made earlier and making it explicit that the information available to private agents at time $t-1$ contains knowledge of $x_{t-\tau}$ for all positive τ, we are led to:

$$x_t^e = \Phi(L)x_{t-1} \tag{259}$$

and finally to:

$$By_t = Ax_t + D(B-D)^{-1}A\Phi(L)x_{t-1} + \eta_t, \tag{260}$$

which involves only observable variables and disturbances. For clarity we may call this equation the "*final backward structural form*", in contrast with the "*initial structural form*" (255).

Thus, we end up with an equation which is quite similar to the structural form of the macroeconometric models discussed in Part 3 of this chapter. The main difference is that the restrictions on the coefficients of lagged values of exogenous variables appear to be more complex than we had envisaged in our discussions of Part 3; but that is not really essential and concerns only the simple-minded representations selected in this book. More important for our present purpose is the fact that, even though we started with (255) from a static model, what comes from the role of expectations, the matrix D, is now entangled with what represents features of the exogenous variables process. As we shall see later, this complication concerns the Lucas critique, tests of parameter invariance and policy evaluation.

2. Let us now introduce in our discussion *current expectations about future values* and, for so doing, just replace (254) and (255) by:

$$y_{t+\tau}^e = E(y_{t+\tau}/J_t) \quad \text{for } \tau = 1, 2, \ldots \tag{261}$$

$$By_t = Dy_{t+1}^e + Ax_t + \eta_t. \tag{262}$$

We can no longer use such a convenient equation as (256) in order to finally express the unobserved vector y^e_{t+1} in terms of observables. But, taking for all future periods mathematical expectations conditional on J_t, we may still derive from (262):

$$By^e_{t+\tau} = Dy^e_{t+\tau+1} + Ax^e_{t+\tau}, \quad \tau > 0. \tag{263}$$

This is a difference equation which, when solved, will give y^e_{t+1} in terms of all the future $x^e_{t+\tau}$. Finally an equation similar to (260) will so be obtained.

However, existence of such a solution requires appropriate properties of the matrices $B^{-1}D$ and $B^{-1}A$ jointly with the process ruling exogenous variables. For simplicity assume this process to be just a first order autoregression:

$$x_t = \Phi x_{t-1} + \xi_t \tag{264}$$

with the characteristic roots of the matrix Φ having all modulus less than 1. The process is then stationary and the expected value $x^e_{t+\tau}$ is given by:

$$x^e_{t+\tau} = \Phi^\tau x_t. \tag{265}$$

Considering precisely the difference equation (263), we are naturally led to think of the solution for y^e_{t+1} as being given by:

$$y^e_{t+1} = Hx_t \tag{266}$$

with

$$H = \sum_{\theta=0}^{\infty} \left[B^{-1}D\right]^\theta B^{-1}A\Phi^{\theta+1}. \tag{267}$$

Two conditions are however required, namely convergence of the sum appearing in the definition of H and convergence to zero of the matrix $(B^{-1}D)^\theta$ when θ increases indefinitely, so that the value of y^e_{t+h+1} for h far distant in the future has no weight in the calculation of y^e_{t+1}. The second convergence requires the characteristic roots of $B^{-1}D$ to have all modulus less than 1 which, added to the assumed property of the roots of Φ, will also imply convergence of (267).

Assuming these properties to hold we are led here to the following model on observables:

$$By_t = (DH + A)x_t + \eta_t \tag{268}$$

which is the final backward structural form for this case. Two remarks must be added.

First, if instead of (264) we had assumed a more general stationary autoregressive process of order h, i.e., Equation (258) with a polynomial of degree

$h - 1$, the expected value $x^e_{t+\tau}$ would have been a linear function of not only x_t but also the $h - 1$ lagged values $x_{t-\tau}$ for $0 < \tau < h$. The model on observables would have involved those lagged values as well; its structure would have been similar to that of (260), obtained in the case in which only past expectations about current values appeared.

Second, Equation (268), or its generalization to a less special model, leads to a well defined solution for the stochastic process $\{y_t\}$. But it is not the only solution. If we want to be precise, we need to qualify it as "the forward solution". The reason for the multiplicity of solutions was explained in Section 1.5 of Chapter 8, where we referred to the simple model (14), corresponding to Equation (262) here. We derived the difference Equation (16) similar to (263) and wrote the forward solution as Equation (17). We then showed how other solutions could be found, each one being derived from the forward solution by perturbations coming from an arbitrary martingale random process. We even envisaged the possibility of a "backward solution".

3. Given the assumptions made above we may consider *the multiplicity of solutions* as a mathematical oddity with no economic significance. The solution implied by (268) is well defined and well corresponds to what we aimed to capture: a forward-looking behaviour, taking account of what is known about the future evolution of exogenous variables, when this knowledge is portrayed by the same stationary stochastic process as generated these variables during the a fairly recent past. We see no reason to pay attention to other solution than this one.

The same sentiment will hold more generally in cases which are not exactly covered by (262)–(264) but either are direct generalizations or lead, may be after a change of variables, to a core system which gives its dynamic properties to the full model and belongs to the same category as (262)–(264).

But what would happen if the properties allowing derivation of (266) from (263) did not hold, for instance if some characteristic root of $B^{-1}D$ had modulus equal to 1 or larger? These would be cases in which forward-looking behaviour would be subject to an embarrassing indeterminacy. Actually, such a possibility is already familiar to us. We found it when discussing deterministic growth models in Chapter 6: processes ruled by a difference equation, such as (263) here, and meant to apply over an indefinite future, starting from a given initial situation, may have a multiplicity of admissible solutions.

We shall not look here at the mathematical theory of linear multivariate rational-expectations models because, except in the favourable case exhibited above, the implications for our present subject would be disappointingly vague, considering the cost of the mathematical investment[229]. We shall rather think in terms of applications.

In any application for which an explicit solution such as (266) is not used (this will be the normal case with non-linear systems), the difference equation such as (263) will be solved for a future period limited to a finite horizon $T = t + h$, given an independently chosen terminal condition, for instance, the value $\hat{y}_{T+1}$ of y^e_{T+1}. There will then be h equations such as (263) for the determination of the h vectors $y^e_{t+\tau}$ $(\tau = 1, 2, \dots, h)$ from the values of the vectors $x^e_{t+\tau}$. The macroeconomist in charge of the application will have to choose the length h of the horizon and the terminal conditions $\hat{y}_{T+1}$.

In order to ascertain whether the horizon is sufficiently long or whether appropriate terminal conditions have been chosen, so that the solution over the period of interest may be confidently retained, numerical sensitivity analysis is needed. Sensitivity to variations in the terminal date indicates how much the solution can be regarded as a genuine product of the model rather than a product of the assumed terminal conditions. Evidence from changes in the type of terminal conditions, together with evidence from the different solution periods, provide an assessment of the degree of confidence to be placed in the predictions obtained in the application.

Only experience can tell when this degree of confidence is likely to be found high, or when in contrast troubles will be revealed. Only experience can build a research strategy for dealing with the latter case. We shall try to say more in a moment about was learned thus far from this experience.

4.2.4. *Policy analysis with linear models*

Moving towards policy analysis we must introduce the vector z_t of policy instruments, the vector x_t concerning now only the "forcing variables" or equivalently the variables on surrounding conditions. This introduction will lead to a more complex modelling, even though the basic ideas are fairly straightforward, given what we have already seen.

The initial system of structural equations will now contain the vector of instruments. For instance, in each one of the two cases examined, (255) and (262), we shall replace Ax_t by $Ax_t + Cz_t$ with a new matrix C. Then the vector z^e_t, defining the values that are expected by private agents, appears in the analysis in parallel with the vector x^e_t, for instance, in (256) or (263). Of

[229] The interested reader may want to study M. Binder and H. Pesaran, Multivariate rational expectations model and macroeconometric modeling: a review and some new results, in: H. Pesaran and M. Wickens (eds.), *Handbook of Applied Econometrics: Macroeconomics* (Blackwell, Oxford, 1995).

course we stick to the assumption of rational expectations. But we cannot just transpose the autoregressive process (258) or (264).

The spirit of the rational-expectations hypothesis requires that private agents know the policy rule which will determine the value of the vector z_t; at least they know it approximately. And this policy rule is unlikely to neglect observation of the endogenous variables, which normally are the subject of concerns on the part of policy makers. The transposition of (258) so requires something like:

$$z_t = \Theta(L)y_{t-1} + \Psi(L)x_{t-1} + \zeta_t, \tag{269}$$

where the shock ζ_t is normally unobserved by private agents, who are assumed here to take it as making a white noise process. Thus, an element of uncertainty remains for agents with respect to what public authorities are exactly deciding, i.e. which intended deviations from the mean policy rule they are selecting.

We shall not try to exhibit the analytical argument incorporating (269) and leading to a final backward structural form comparable to (260) or (268). The logic is exactly the same as for the two cases previously discussed. We gathered now enough elements for reaching a few conclusions which have a general validity.

First, the final backward structural form gives us the relation between observable variables, a relation that we can fit on observed time series for estimation of the parameters. But this form relies on all the assumptions made in order to reach it, not only the initial structural form, such as (262) together with the expression (261) given to the rational expectation hypothesis, but also a specific assumption about the content of the information set J_t, for instance, here: (i) that this set contains observation of all the exogenous vectors $x_{t-\tau}$ and $z_{t-\tau}$ up to x_t and z_t, as well as observation of all the endogenous vectors $y_{t-\tau}$ up to y_{t-1}; (ii) that it contains knowledge of the full true model in its quantitative form, not only the initial form (262) but also the process ruling exogenous variables on surrounding conditions, such as (264), and the policy rule, such as (269). Within this strong assumption about the information set of private agents is embodied *an assumption of structural stability*: the process ruling surrounding conditions is stable (it is believed to remain unchanged from the past to the future), and the policy rule is stable (it is also believed to remain unchanged).

Second, we can now read again the argument given by Lucas in its critique of macroeconometric policy evaluation; we shall so understand the meaning of Lucas' critique. Starting with the last step 4 in the argument displayed earlier here and moving backward, we may say: 4 – indeed by definition, a change in the policy *rule* is an alteration in the process (269) generating the instrument variables; 3 – indeed, the formation of private expectations is likely to be revised when private agents are conscious that the structure of the process generating the policy decisions is altered (in the same way as, according to (265) in the second case discussed above, the formation of expectations about $x_{t+\tau}$ would be revised if agents were conscious of a change in the matrix Φ); 2 – indeed, in the same case the value given to the matrix H would probably have to be revised; 1 – indeed, again in the same case and more generally when the formation of expectations is revised, the quantified final-backward-structural form is very likely to also be revised. We now understand the exact meaning of Lucas' "syllogism", and this also permits us to evaluate its force in concrete situations.

Third, we also understand K. Wallis when he explains in his 1980 article that "a practical response" to Lucas' critique can "be devised, by retaining the notion of an economic structure (incorporating the rational expectations hypothesis) that is invariant to the structure of exogenous processes" (p. 50). More precisely the practical response is to estimate what we call here the initial structural form. This estimation can in principle be derived from estimation of the final backward structural form.

There are of course a number of conditions, in the first place that everything be correctly specified: the initial structural form, for instance (262) with $Ax_t + Cz$ replacing Ax_t, the process of the surrounding conditions, (258) say, the process of the policy rule, (269) say. It is also necessary that all the parameters appearing in those specifications remained unchanged during the sample period. It is finally necessary that the parameters of the initial structural form be identifiable from the estimate of the final backward structural form. Actually, this identifiability, which is discussed in Wallis'article depends on whether there are enough prior restrictions on the matrices appearing in the specification. Identifiability is unlikely to be a problem in simple specifications.

Once the initial structural form, the process of surrounding conditions and the past policy rule have all three been estimated, policy evaluations responding to the Lucas' critique have simply to be clear on what is assumed in each variant of the policy examined. It will commonly be acceptable for all variants to maintain unchanged as estimated the initial structural form and the process on surrounding conditions, but to use different sets of values for the parameters appearing in the policy rule, specially in (269) for those of the matrix $\Theta(L)$ characterizing the reaction of instruments to changes in endogenous variables, such as employment or the rate of inflation.

Fourth, study of the effects of perceived and credible major changes in the policy rule does not cover the whole field of policy analysis. We very often need to know what "normal policy-making" achieves. We need to know better what the values of the many relevant "multipliers" may be. Readers of this book realize that the days of improvement in this basic knowledge are not yet bygones. Effects of normal policy-making, as against those of sudden and permanent shifts in the policy regime, can and will be usefully analysed within the family of conventional structural macroeconometric models, even though these models may be viewed as providing just final backward structural forms.

But, fifth, models with forward expectations allow a distinction to be drawn between anticipated and unanticipated policy changes, or between changes expected to be temporary and those expected to be permanent; they thus allow to study the effects of policy announcements. This is an important addition to what can be done with conventional macroeconometric models, which may be said to be purely backward-looking.

What is involved in the anticipated-unanticipated distinction, or in the temporary-permanent distinction, or still in policy announcements, concerns the content of the information set. When a full rational-expectations model is displayed, we see exactly where the content of J_t matters in the computation of the effects of a policy variant, therefore also in the computation of the difference in the results of two variants which differ only by the policy announcement, say. We may compute in particular the immediate effect in period t of the announcement of a policy to be implemented from period $t+1$ on. And so on[230].

Sixth, we are now in a position to understand the nature of that econometric test of the Lucas' critique which was performed on a particular backward-looking econometric equation by Favero and Hendry (op. cit.). This equation, described in this book at the end of Section 4.6 of Chapter 2, concerns the US demand for money and was fitted on quarterly data over almost thirty years[231]. This might look like a good candidate for showing the worth of the Lucas' critique because the equation contains such regressors as the rate of inflation, interest rates, or the volatility of the yield of bonds, for which purely adaptive expectations are questionable, given the fact that a number of important changes in monetary policy occurred during the period. We may think that adaptive

[230] For a simple analogy showing the point of this paragraph, let us go back to the model defined by (262)–(264) and compare the solution given by (268) with what would result from the knowledge that a permanent shift of size a in $x_{t+\tau}$ would occur from period $t+1$ ($\tau = 1$) on: the constant vector a would be added in the right-hand side of (264) for period $t+1$, whereas $x_{t+\tau}$ and $x_{t+\tau-1}$ would be replaced in (264) for later periods ($\tau > 1$) respectively by $(x_{t+\tau} - a)$ and $(x_{t+\tau-1} - a)$. Then a would be added in the right-hand side of (265) for all positive τ. The difference Equation (263) shows that the constant vector $(B - D)^{-1}\,Aa$ would have to be added to all $y^e_{t+\tau}$, hence to y^e_{t+1}. The impact on y_t would result from (262).

[231] Y. Baba, D. Hendry and R. Starr, The demand for M_1 in the USA, 1960–1988, *Review of Economic Studies* (January 1992).

expectations performed particularly badly when deviations of the regressors from equations explaining their respective movements were abnormally large or variable. We may then suspect that introducing the residuals from those equations, or/and three period moving standard deviations of these residuals, as additional regressors in the demand-for-money equation would improve the fit and so reveal the weakness of the backward-looking adaptive expectations. Actually Favero and Hendry, who ran a large number of regressions with different sets of such additional regressors, testing in each case whether their addition brought a significant improvement, did not find any sign of confirmation of the Lucas' critique.

4.2.5. *A monetary policy rule for an economy dominated by rational expectations*

Once it has been understood, the logic of the rational-expectations approach may be used to deal with traditional macroeconomic problems as they would appear in an economy in which private expectations would conform with the predictions of the approach. We are going to so deal with problems initially posed here in Parts 3 and 4 of Chapter 8. We shall then draw from a recent paper of R. Clarida, J. Gali and M. Gertler[232].

Our discussion of monetary policy rules began in Section 3.1 with the simple framework used by R. Barro and D. Gordon in their famous 1983 model. The behaviour of the private economy and of the shocks to which it was exposed, Equations (81) and (82) then, appeared at first sight consistent with current ideas on price and quantity adjustments, as they were formalized in particular for aggregate demand analysis. But closer examination showed that the hypothesis of a direct control of the rate of inflation by the central bank twisted these ideas. Simultaneously, introduction of rational expectations as concerning only the rate of inflation may have looked too special, as it concerned only policy and not endogenous variables of the model.

Clarida et al. take the inflation rate g_t as being endogenous besides the "output gap", which we denote here q_t and is meant to be the difference between the logarithms of respectively current output and capacity (or equivalently the "natural level of output"). Their model is then written as:

$$q_t = q^e_{t+1} - \varphi[i_t - E_t g_{t+1}] + \eta_{qt}, \tag{270}$$

$$g_t = \beta g^e_{t+1} + \lambda q_t + \eta_{gt}, \tag{271}$$

$$\eta_{qt} = \mu\eta_{q,t-1} + \zeta_{qt}, \qquad \eta_{gt} = \rho\eta_{g,t-1} + \zeta_{gt}. \tag{272}$$

[232] R. Clarida, J. Gali and M. Gertler, The science of monetary policy: a new Keynesian perspective, *Journal of Economic Literature* (December 1999).

The first equation explains how output is determined as a function of the expectation of its level in the following period, of the expected real interest rate and of a random shock η_{qt}. The second equation explains how the rate of inflation is determined from its expectation in the next period, the degree of capacity utilization and a shock η_{gt}. The two last equations assume that the shocks follow independent first-order autoregressive processes, $\{\zeta_{qt}\}$ and $\{\zeta_{gt}\}$ being pure white noises. The parameters φ, β, λ, μ and ρ are meant to be positive, β being the discount factor.

The authors explain why they chose this representation of the private economy, η_{qt} and η_{gt} being respectively demand and price shocks. Equation (270) differs from a traditional IS curve by the presence of expected future output, which may be explained for instance by consumption smoothing. The Phillips curve (271) may be explained by staggered price setting and monopolistic competition. Neither output nor inflation depends on past economic conditions; they depend only on present and expected future conditions (in Part 6 of their paper Clarida et al. show how their results ought to be revised if past conditions would also influence the current equilibrium). We see that the system belongs to the same family as (262) posed in Subsection 4.2.3 above, except that it involves no variable on surrounding conditions (x_t then) but rather a term like Cz_t in Subsection 4.2.4, involving the policy instrument (z_t then, the nominal interest rate i_t now). So we may write:

$$By_t = Dy^e_{t+1} + Ci_t + \eta_t \tag{273}$$

with

$$\begin{aligned} y_t &= \begin{bmatrix} q_t \\ g_t \end{bmatrix}, & B &= \begin{bmatrix} 1 & 0 \\ -\lambda & 1 \end{bmatrix}, \\ D &= \begin{bmatrix} 1 & \varphi \\ 0 & \beta \end{bmatrix}, & C &= \begin{bmatrix} -\varphi \\ 0 \end{bmatrix}. \end{aligned} \tag{274}$$

The authors discuss the appropriate policy rule for the choice of i_t in this economy. We shall not reproduce in full their discussion, which would amount to rewrite Part 3 of Chapter 8 with a better model of the economy and a better representation of the policy instrument (i_t rather than g_t). We shall indeed limit attention to the choice of a parameter ω to be applied in a fully credible policy rule that would target the rate of capacity utilization (on the concept of a target rule, remember Section 4.4 of Chapter 8). The choice of ω will, of course, reflect the aim of the monetary authority and the respective importance it is giving to the stimulation of economic activity and to the restraint of inflation.

Like Barro and Gordon the authors define a loss function to be minimized, namely:

$$L = E_t\left\{\sum_{\tau=0}^{\infty}\beta^{\tau}\left[\alpha q_{t+\tau}^2 + g_{t+\tau}^2\right]\right\}, \tag{275}$$

α being all the higher as more consideration is given to stimulation of activity (introducing in L deviations such as $g_{t+\tau} - \hat{g}$ instead of $g_{t+\tau}$ would not mean much change in the following analysis). To complete the problem of the monetary authority we must still be precise about the information available to this authority. Clarida et al. assume that, when choosing i_t, the authority knows η_{qt} and η_{gt}: thanks to its constant scrutiny it is immediately informed of current economic conditions. Moreover it pays particular attention to the price shock: its target q_t^* is all the lower as the shock is higher:

$$q_t^* = -\omega\eta_{gt}. \tag{276}$$

Actually the choice between a target on output or one on inflation is immaterial because, given the price behaviour of the economy, (276) is equivalent to:

$$g_t^* = \frac{1-\lambda\omega}{1-\beta\rho}\eta_{gt}. \tag{277}$$

This follows from (271) and the second of Equations (272). Clearly the best choice of ω then is the one minimizing $\alpha q_t^{*2} + g_t^{*2}$, hence:

$$\hat{\omega} = \lambda\nu, \quad \nu = \left[\alpha(1-\beta\rho)^2 + \lambda^2\right]^{-1}. \tag{278}$$

Given this choice of ω, (276) and (277) lead to:

$$q_t^* = -\lambda\nu\eta_{gt}, \qquad g_t^* = \alpha(1-\beta\rho)\nu\eta_{gt}. \tag{279}$$

It is now obvious from (270) that the instrument rule fixing the nominal interest rate i_t follows from the target rule through:

$$\varphi\left[i_t - \alpha(1-\beta\rho)\nu\rho\eta_{gt}\right] = \lambda\nu(1-\rho)\eta_{gt} + \eta_{qt} \tag{280}$$

or

$$i_t = \gamma g_{t+1}^e + \frac{1}{\varphi}\eta_{qt} \tag{281}$$

with

$$\gamma = 1 + \frac{\lambda(1-\rho)}{\varphi\alpha(1-\beta\rho)\rho}.$$

The instrument rule means that the *real* interest rate $i_t - g^e_{t+1}$ increases as a function of both the output and the price shocks.

Overall, such results make sense and convey the same qualitative message as the one following from our study of monetary policy under adaptive expectations in Part 4 of Chapter 8. In their exact formulas they, of course, depend both on the model of the economy, i.e. on Equations (270) to (272), and on the specification of the policy problem. Full credibility of the rule implies commitment of the monetary authority to follow the rule. Limitation to the simple target (276), means that the resulting level of capacity utilization will be independent of the demand shock. This suggests that the policy rule would not be found optimal within a wider set of rules, a point recognized and discussed by Clarida et al.

Within the general scope of this section the analysis brings an interesting conclusion of a different kind, namely that the multiplicity of solutions of linear models with rational expectations may be in practice less troublesome than our discussion in Subsections 4.2.3 and 4.2.4 might have suggested. Let us look at this point a little more closely, comparing with hindsight the approaches of this subsection to the one of the two earlier subsections.

We accepted, in the argument we just gave, the hypothesis that, since the monetary authority had the target rule implying (276) and (277), private expectations of $q_{t+\tau}$ and $g_{t+\tau}$ would coincide with what expectations of $\eta_{g,t+\tau}$ would imply for $q^*_{t+\tau}$ and $g^*_{t+\tau}$. Now, assume that private agents know only the instrument rule (281) and believe it will be exactly followed in the future. What will the approach of Subsections 4.2.3 and 4.2.4 predict?

Taking (281) into account we now replace (273) by:

$$By_t = D^* y^e_{t+1} + K\eta_t \tag{282}$$

with

$$D^* = \begin{bmatrix} 1 & -\varphi(\gamma - 1) \\ 0 & \beta \end{bmatrix}, \qquad K = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}, \tag{283}$$

or

$$y_t = Ry^e_{t+1} + K\eta_t \tag{284}$$

with

$$R = B^{-1}D^* = \begin{bmatrix} 1 & -\varphi(\gamma - 1) \\ \lambda & \beta - \lambda\varphi(\gamma - 1) \end{bmatrix}. \tag{285}$$

Taking expectations at time t on both sides of (284) written for $t + \tau$ instead of t, as we did when going from (262) to (263), we find:

$$y^e_{t+\tau} = Ry^e_{t+\tau+1} + \rho^\tau K\eta_t. \tag{286}$$

Iterating up to horizon $t+h$ and reporting y^e_{t+1} so found in (284), we find:

$$y_t = R^h y^e_{t+h} + \sum_{\tau=0}^{h-1} (\rho R)^\tau \cdot K\eta_t. \tag{287}$$

What will happen at the limit, for h increasing to infinity, clearly depends on the characteristic roots of the matrices R and ρR. The sum of the characteristic roots of R is equal to $1+\beta-\lambda\varphi(\gamma-1)$ and their product is equal to β, a positive number smaller than, but close to, 1.

Except under quite unlikely values of the parameters (λ too small or $\alpha\varphi\rho$ too large), the two characteristic roots of R have modulus smaller than 1. *A fortiori* this is the case also for the characteristic roots of ρR. Hence R^h tends to the zero matrix and the sum in (287) tends to the matrix $(I-\rho R)^{-1}$.

We thus expect that (282) has the solution given by:

$$y_t = (I-\rho R)^{-1} K\eta_t. \tag{288}$$

Indeed, this implies for any h:

$$y^e_{t+h} = \rho^h (I-\rho R)^{-1} K\eta_t. \tag{289}$$

Used jointly these two values of respectively y_t and y^e_{t+h} trivially fulfil (287)[233].

If is left to the reader to check that, given (278), Equation (288) leads exactly to $q_t = q^*_t$ and $g_t = g^*_t$. The constructive approach of this subsection so agrees with the natural forward solution to which leads, in this case, the more general approach of the two preceding subsections. Other solutions to (287) written for all values of h also exist, but how could they be the focus of attention for rational agents, when the simpler and more natural one given by (288) is available?

4.3. Macroeconometric models with rational expectations in practice

The idea to introduce in macroeconometric models latent variables representing private expectations and to assume that these expectations are model-consistent has some impact on the present practice. Experience gained in this

233 Note that (288), hence (289), jointly fulfil (287) as soon as $I-\rho R$ has an inverse, no matter what are otherwise the characteristic roots of R. But if some of these roots have modulus larger that one, (288) will define an unstable solution, in the sense that some small deviations of y^e_{t+h} from the right-hand side of (289) will be magnified by R^h in the determination of y_t by (287).

respect over the last twenty years opens new possibilities to model builders and helps to defuse the Lucas' critique.

Readers of this book should like to find here a perfect account of this experience. The author is not best placed to provide such an account but will have a try, giving a few references and taking advantage in particular of the book of Paul Fisher[234] where the UK experience appears. He will first report what he knows about the history of this development of his discipline, then write about the technical problems, further raise the question of knowing the properties of macroeconometric models in which rational expectations play an important role, finally report the experience of the team of econometricians working at the US Federal Reserve Board.

4.3.1. *An overview*

R. Fair[235] is the author of the first experiment with incorporating model-consistent forward expectations in econometric policy analysis. Three equations of an earlier medium sized model of the US economy are revised so as to contain rational expectations of interest rates and stock prices. The purpose is then not to test the rational-expectations hypothesis but to examine the sensitivity of the properties of the model to incorporation of this hypothesis. Particular attention is given to the estimated effects of an unanticipated permanent increase in government consumption. In the earlier model with adaptive expectations this had a gradual effect over time on interest rates. In the new model knowledge that the short-term rate would increase in the future is reflected immediately into long-term rates. This leads to a lower initial increase in real GNP, but the difference is not so large: it culminates in the fourth quarter after the increase in government consumption (multipliers respectively equal to 1.3 and 0.9). After considering effects on other variables and effects of anticipated increases, the author concludes: "The rational expectations assumption is clearly of some quantitative importance... especially with respect to timing and anticipated changes".

In his econometric modelling of the then called "new classical approach" to supply-side economics, P. Minford set up, again around 1980, the "Liverpool model", the first in the UK to allow for rational expectations[236]. In its 1985

[234] P. Fisher, *Rational Expectations in Macroeconometric Models* (Kluwer Academic Publishers, Dordrecht, 1992).

[235] R. Fair, An analysis of a macroeconometric model with rational expectations in the bond and stock markets, *American Economic Review* **69** (1979) 539–52. The model is also presented in Chapter 11 of Fair's 1984 book (op. cit.).

[236] P. Minford, S. Marwaha, K. Matthews and A. Sprague, The Liverpool macroeconomic model of the United Kingdom, *Economic Modelling* **1** (1984) 24–62.

version the model has 34 endogenous variables and 6 forward-expectations terms (concerning capacity utilization, real interests on the public debt, the real exchange rate – one and five years ahead, the inflation rate – over one and five years ahead). As we shall see later a problem appears with this model, namely that simulations lead to solution paths which are sensitive to ancillary assumptions made for definiteness of the computations.

Shortly after, rational expectations were adopted in the standard operating mode of forecasting and policy analysis for two much larger models of the UK economy. The London Business School model, which has 1258 endogenous variables, contains 3 expectational terms about stock prices (for gilt-edged stocks, equities and overseas assets respectively). The model of the National Institute of Economic and Social Research, which has 199 endogenous variables, contains 40 expectational terms in 11 variables with leads of up to four periods ahead (real GDP, three sectoral output index, consumer prices, wholesale prices, average earnings, personal disposal income, interest rate, effective exchange rate, rate of employer's national insurance contributions).

In the late 1980s macroeconometric models incorporating rational expectations were built for a few other countries, and also at the multicountry level. Reference to them are given in an article reporting interesting empirical evidence coming from a Dutch model[237]. More recently still such models were developed in the US. For example, J. Taylor (1997, op. cit.) reports how a rational-expectations model used at the Federal Reserve Board is able to estimate the effects on interest rates of a multiyear plan to reduce the future budget deficit. The same model can explain how the plan will have a smaller short-run contractionary effect if it is credible[238].

The conventional macroeconometric model of the Dutch Central Bank was reexamined during the second half of the 1980s (see J. Bikker et al., op. cit.). Nine equations in which rational-expectations variables might occur were studied in detail. Of these, seven were reestimated with success, i.e. including one or more rational-expectations variables. These equations concerned private consumption, the average wage bill per employee, investment in dwellings, the exchange rate, the short-term interest rate, a stock price index, the long-term funds raised by business. The changes with respect to the earlier estimated equations were small. Even the coefficients of the rational-expectations variables did not differ markedly from their adaptive predecessors. From an exten-

[237] J. Bikker, P. Van Els and M. Hemerijck, Rational expectation variables in macroeconomic models – Empirical evidence for the Netherlands and other countries, *Economic Modelling* (July 1993).

[238] F. Brayton, A. Levin, R. Tyron and J. Williams, The evolution of macro models at the Federal Reserve Board, *Carnegie Rochester Series on Public Policy* (1997).

sive study of variants delivered by the revised model the general conclusion was drawn that the most important differences between the impulse effects in the adaptive-expectations and rational-expectations versions turned up mainly in the short run, because "rational" agents were assumed to quickly incorporate new information in their expectations and were thus reacting faster to change. On the other hand, the results for the medium term of four years generally showed great similarities.

In their brief survey of other models J. Bikker et al. show that, up to 1993, rational-expectations variables had been systematically introduced for only few variables: the inflation rate, the exchange rate and the interest rates. Even such a variable as rational expectations about households' disposable income appears in fewer than half of the rational-expectations models discussed.

4.3.2. *Some "technical" issues*

(i) Subsection 4.2.3 above will help us to now understand how the presence of model-consistent expectations makes the use of macroeconometric models more complex. We already knew that those models were in practice non-linear and that, in forecasting or policy analysis, solution of the models for the required simulations had to be numerically computed. Given the complexity of the model, the computational procedure had to be iterative, each iteration aiming at a closer fit to the system of equations than the previous iteration had achieved.

In conventional backward-looking models the procedure could work sequentially, with iterations concerning future periods in succession, one by one. Temporary equilibria were so determined, each from knowledge of its predecessors, no feedback being required. The presence of forward expectations imposes in contrast a simultaneous determination of all temporary equilibria up to the horizon of the simulation. In other words, the iterations in the computation must work on full *equilibrium paths* rather than on temporary equilibria. We shall not describe here the computational procedures which have been found workable and efficient, but we understand that they may be heavy to carry.

From our examination of linear models we remember that, in models with unbounded horizons, there exists usually a multiplicity of consistent solution paths. Often only one of these paths is "long-run stable" in a meaningful sense, attracts all other paths and therefore appears to be relevant for the characterization of long-run properties. But, even in such a favourable case, computations cannot consider an unbounded future: their objective is to find the, usually unique, finite equilibrium path meeting some terminal conditions. The choice of these terminal conditions may be innocuous or not, depending on their precise specification, on the length of the horizon and on the characteristics of the model. Let us look a little more closely at this choice.

In the favourable case, when we know or can safely assume that all indefinite paths would converge to a particular stable path (we call it for simplicity "the stable path"), terminal conditions have to be selected in such a way as to ensure that the solution would approximate the stable path up to the horizon. If the nature of this path is known *a priori*, for example, from analysis of the "steady-state" properties of the model, the terminal conditions must incorporate this knowledge. It is seldom precise enough to indicate even rough approximations of the values of the endogenous variables at the horizon. It is usually more qualitative. Typically it will show that, after some time, some endogenous variables will be about constant, the others growing at an about constant rate. In order to reflect this knowledge, the terminal conditions will then say that $y_{j,T+1}$ will have to be equal to y_{jT} in the solution for a variable j of the first group, and that the rate of growth from y_{jT} to $y_{j,T+1}$ will have to be equal to that from $y_{j,T-1}$ to y_{jT} for a variable j of the second group.

The validity of such terminal conditions may require, however, a relatively long solution period so as to ensure that the solution is not substantially distorted by the assumption that, at the end of this period but starting from the given initial conditions, the path will have reached the stable course. Moreover the constant-level or constant-growth-rate terminal conditions may create difficulty in the computations. Whereas existing computational procedures are defined in such a way as to accommodate these conditions as well as those specifying given terminal values, convergence of the successive iterations towards a well defined path is achieved within a smaller set of models; non-convergence, if it occurs, imposes to the user to shift to a different set of terminal conditions, in practice to given-values conditions.

A *fortiori* the choice of terminal conditions is delicate when prior analysis of the properties of the model does not lead to the full determination of sufficiently binding long-run constraints. As we concluded at the end of Subsection 4.2.3, only experience can build a research strategy for dealing with such unfavourable cases.

(ii) The book of P. Fisher (op. cit.) reports what this experience has been with the three UK models mentioned earlier. But before doing so, it studies in its Sections 4.2 and 4.3 the role of terminal conditions for a very simple linear model with just one endogenous variable p_t (the price level in the literature about this model) and one equation:

$$p_t = \alpha E(p_{t+1}/p_{t-1}) + u_t \tag{290}$$

the process u_t being exogenous and α a known number. For this "demonstration model" various terminal conditions are considered in order to determine the solution $\{\hat{p}_{t+\tau}\}$ for $\tau = 0, 1, \ldots, T - t$, in particular:

- either a *fixed value* $\hat{p}_{T+1} = \bar{p}_{T+1}$ where $\bar{p}_{T+1}$ is set from off-model analysis,
- or a *constant level* $\hat{p}_{T+1} = \hat{p}_T$
- or still a *constant growth rate* $\hat{p}_{T+1} = \hat{p}_T^2 / \hat{p}_{T-1}$.

What will result from the choice will depend on the value of α and on the characteristics of the process followed by u_t.

Fisher considers a few simple cases. In the first one $|\alpha| < 1$ and

$$E_t u_{t+\tau} = u^* \quad \text{for } t + \tau > T, \tag{291}$$

where E_t means the mathematical expectation at time t and u^* is a constant number. Since $|\alpha| < 1$ we are in the favourable case of a unique stable forward solution:

$$p_t = \sum_{\tau=0}^{\infty} \alpha^\tau E_t u_{t+\tau} \tag{292}$$

implying in particular $E_t p_{t+\tau} = p^* = u^*/(1 - \alpha)$ for $t + \tau > T$.

In a second case, the number α has again absolute value smaller than 1, but (291) is replaced by the hypothesis that u_t follows an autoregressive process implying:

$$E_{t+\tau} u_{t+\tau+1} = \gamma_0 + \gamma_1 u_{t+\tau}. \tag{293}$$

We are still in the favourable case of a unique stable forward solution given by (292) implying convergence of $E_t p_{t+\tau}$ to $p^* = \gamma_0/(1 - \alpha)(1 - \gamma_1)$ for τ increasing to infinity.

Fisher has no difficulty in showing that, in these two cases, no matter whether the terminal condition stipulates a fixed value, a constant level or a constant growth rate, the solution path up to the finite horizon T would converge to the stable solution path if T would increase to infinity. With a finite T an error remains, whose significance can be judged, in the demonstration model, by its size in the terminal period.

This size is computed in the second case defined above for various values of the triple (α, γ_1, T). The following qualitative results emerge:

- the fixed value $\bar{p}_{T+1} = p^*$ is the most effective in minimizing the terminal period error only when α is near unity and γ_1 not large (something like $\alpha = 0.95$ and $\gamma_1 = 0.6$),
- the constant growth rate terminal condition is relatively effective only when γ_1 is near unity and α not large ($\gamma_1 = 0.95$ and $\alpha = 0.6$, say),
- otherwise, the constant level condition is the most effective.

(ii) Fisher further asks: what should econometricians do with the demonstration model (290) when α happens to be larger than 1 in absolute value? Trying to answer the question is important because the case captures a problem that occurs more generally, although perhaps in milder forms: it occurs as soon as the model has not one and only one stable solution. But no clear answer seems to exist.

We appreciate the difficulty when learning that, according to a precise mathematical analysis applying when u_t follows a stationary process, there is a multiplicity of stable solutions[239], with among them the backward solution exhibited here in Section 1.5 of Chapter 8. But, as will be presently suggested, none of these solutions seems to provide an economically meaningful reference for evaluating the "error" in the solutions computed over a finite period.

Moreover Fisher argues that another problem occurs with the existing computational procedures when they are applied with the constant-level terminal condition (the same problem occurring with the constant-growth-rate condition): the numerical iterations will not converge, whereas the finite horizon equilibrium path meeting the terminal condition in question exists and can be found from (290) by a purely analytical method. Fisher concludes: "If fixed values are used one may therefore select a solution to this model even when there is more than one stable solution and this non-uniqueness will not be apparent. Hence the use of such conditions may be misleading".

The author of this book thinks that the problem might well be deeper, because macroeconomists could hardly be pleased if they were provided even with the exact bunch of stable solutions starting from given initial conditions. Let us consider simply the backward solution, in order to avoid speaking about differences of martingales. Would any macroeconomist, concerned with knowing the impact of agents' expectations, consider that this backward solution makes economic sense? Would he or she be more interested after learning that the stable solution is assumed to have been in place over an indefinite past, ruled all the way by the same process (270) with the same value of α?

Last but not the least, and speaking more generally, as soon as we theorists begin to hesitate about what may be the proper selection of the appropriate future growth path, should not we still more hesitate about the worth of the rational-expectations hypothesis in the case discussed? How could normally rational agents go through the process of finding something like their model-consistent and mutually-consistent expectations?

[239] See M. Binder and H. Pesaran (1995, op. cit.).

4.3.3. *A closer look at the British experience*

We are now well equipped for a closer look at the three UK models studied in Fisher's book, which was born at Warwick within the Macroeconomic Modelling Bureau. We had in Section 3.5 another opportunity to benefit from the careful and deep investigations of the Bureau into the properties of macroeconometric models. We probably could not find a better reference for our present purpose, which is to better understand how models of a particular kind work, both in forecasting and in policy simulations.

In order to complement what was explained in Section 4.2 and also to stimulate the interest of the reader we shall begin with a survey of some results obtained thanks to the models, before turning to sensitivity analysis, the use of terminal conditions and other related difficulties. We shall moreover treat apart the Liverpool model which is meant to offer a different view of the macroeconomy from the one conveyed by most structural models (it is also an annual model whereas the other two are quarterly).

(i) In his Section 5.4 Fisher presents simulations illustrating the difference in effects between anticipated and unanticipated shocks (conventional macroeconometric models ignore the distinction and therefore give the same prediction if the two shocks are otherwise identical). A step increase in real general government expenditure is introduced in quarter 9 keeping fixed nominal interest rates (government finances the increase by an extra issue of money). Two solutions are computed (actually three in the book, but we simplify). The first assumes that the shock is announced and expected since the end of quarter 0 (the start of the simulation), the second solution assumes that the shock becomes known only at the end of quarter 9 and is so completely unanticipated. The book graphs the trajectories of the impacts on GDP and on the nominal exchange rate, which happens to be a crucial variable for understanding the properties of the models (see below).

The models of respectively the London Business School (LBS, 3 expectations terms) and the National Institute of Economic and Social Research (NIESR, 40 expectations terms) give qualitatively identical results: after quarter 9 and up to the end of the graphs (quarter 34 or 23, respectively) the impacts of anticipated and unanticipated shocks are practically the same (a significantly higher GDP, a persistently growing negative effect on the exchange rate, showing a higher inflation rate); as soon as the shock is anticipated, the negative effect on the exchange rate begins to build up; in case of anticipation there is also an early positive impact on GDP, but it remains quite small before quarter 9. The only major difference between the two models is that the LBS model is more responsive overall, particularly for the exchange rate.

In his Section 5.6 Fisher presents simulations illustrating the difference in effects between shocks that are respectively permanent and temporary, rightly believed to be so, even though unanticipated on introduction. The temporary shock, on government expenditure again, is maintained for five or three years, respectively with the LBS or the NIESR model (again conventional models give the same prediction in the first phase, whether the shock is going to be permanent or only temporary). The simulation responses of both models are in general smaller when the shock is expected to be temporary rather than permanent, although the differences are not that substantial for the LBS model. For both models, after the end of the temporary shock, the GDP trajectory returns close to its base level whereas the nominal exchange rate returns to a trajectory parallel to its base.

(ii) The technical choices made for, or the difficulties experienced in, deriving the solution paths are described in Section 4.4 of the book. The author reports the terminal conditions used for the expectations terms and the results of a sensitivity analysis, which considers different lengths of the horizon chosen in the simulations or alternative terminal conditions. His section also contains a few advices to the users of the models.

Owners of the LBS model chose a constant-level terminal condition for the price of gilt-edged stocks and a constant-growth-rate condition for the two other stock prices of their model. Fisher experimented with five alternative sets of conditions: constant-level for the three stock prices, constant-growth-rate for the three, consideration in the three cases of a constancy applied to two quarters a year apart, instead of constancy applied to the two last quarters ("seasonalized-constant-level", say). The book gives graphs of the trajectories of the gilt price found in the sensitivity analysis with horizon-lengths varying from six to nine years.

Since the trajectories show during the three first years very little sensitivity to the technical choices made but may later strongly diverge, Fisher concludes: "the LBS model does seem to possess a [unique stable] path". But it appears, from the results with conditions chosen by the owners of the model as well as with the two other unseasonalized sets of conditions, that the trajectory much depends on which quarter of the year the simulation takes as an end point (this is magnified, particularly for the constant-growth-rate conditions, by a fortuitous turning point of the series at the end point chosen by the model owners). This seasonality is precisely the reason that let Fisher to experiment with seasonalized terminal conditions, which indeed turn out the give a lower influence to the exact choice of the end point and are also found to reduce to practically nothing the sensitivity of the policy simulation responses with respect to this choice.

Overall we may say that model-consistent expectations seem to be easily managed in the LBS model and that a useful suggestion to owners of quarterly models is made as to the choice of terminal conditions. Easiness in the introduction of rational expectations is probably typical of cases where they concern only a limited segment of the model. The situation is not as easy with the NIESR model.

As we saw, the model contains 40 expectations terms in 11 variables. By far the most important of these variables is the real exchange rate. The terminal conditions impose constant growth rates at the horizon to the other 10 variables. Owners of the model imposed at the time to the real exchange rate a terminal value equal to the initial value of this rate after a correction proportional to the relative change, over the full period, in the ratio of exports to imports. Fisher had some difficulty with the application of this terminal condition, which he explains. But the main point is the high sensitivity of the solution to the choice of the horizon, a sensitivity appearing since the first quarters of the simulation and concerning all main variables.

This is explained by the fact that, in the version of the model used by the Institute when the sensitivity analysis started, the dynamic linear real exchange rate equation: (i) was autonomous, containing only the present, past and expected future exchange rate, together with an exogenous variable of the full model, (ii) had an imposed unit root (the coefficients of the exchange rate terms added to zero in the specification). As a result of these two features the full model behaved like a linear dynamic model with a unit characteristic root: long-run evolutions were so subject to an embarrassing indeterminacy to which we referred toward the end of Subsection 4.2.3.

Later versions of the NIESR model, starting in Autumn 1986, incorporate two corrections. The exchange rate equation includes an endogenous current account term, so that the whole system no longer has the built-in unit root. A constant growth rate is used as terminal condition for the exchange rate. However, probably due to a near-unit root remaining, sensitivity to the choice of the horizon is still high so that a very long computation period must be chosen for a stabilization of the simulation results (up from 23 to 40 periods).

These difficulties prompted the Warwick Bureau to launch a specific research programme about what are more generally the interactions between the exchange-rate equation and the rest of the model within large-scale macroeconometric models containing rational expectations about the exchange rate. We shall now extensively quote the introduction of the article in which results of this research are reported[240].

[240] P. Fisher, D. Turner and K. Wallis, Forward unit root exchange-rate dynamics and the properties

"In models of small open economies operating under a floating exchange-rate regime the treatment of the exchange rate is often found to be a major source of cross-model differences... In simulation analysis of several UK models, standardizing on an [exchange-rate equation of the uncovered-interest-parity form where the coefficient of the expected exchange rate is one] reduces differences in their overall properties but does not eliminate them, since differences remain in other channels of the exchange-rate transmission mechanism... This paper presents a simple model reduction technique that provides a convenient means of understanding the behaviour of the exchange rate in policy simulations with large-scale macroeconometric models... The large-scale non-linear model can be reduced to a two-equation system comprising an exchange-rate equation and a feedback [linear] equation. The sensitivity of the reduced model's properties to the dynamic specifications of the exchange-rate equation [and the feedback equation] is assessed; in particular it is readily seen that the long-run equilibrium is independent of the degree to which the [exchange-rate] equation is forward or backward looking... [It is further shown that] information provided by the reduced model... can improve the efficiency of... the numerical solution methods that are used in non-linear forward expectations models".

(iii) Simulations made on the Liverpool model lead to quite different results from those run on the two quarterly models. An unanticipated money-financed step increase in government expenditure brings about just a temporary and relatively small increase in GDP. If anticipated the same shock has a stronger effect, which mainly comes before the increase in government expenditure and is actually a decline in GDP (anticipated future inflation has a negative impact on current GDP). A stronger positive impact on GDP is obtained if the initially-unanticipated shock is expected to be temporary, even for several years, than if it is then expected to be permanent (again this is due to the depressing effect of anticipated inflation, which is more dramatic when the shock is expected to be permanent). Fisher comments: "These conclusions of the Liverpool model suggest that a government expenditure increase will only have a substantial impact on the real economy if it is perceived to be temporary, as in the case of a pre-election boom".

In order to look at the extensive sensitivity analysis made by Fisher on this model, let us first note that, in the structural form of the model, no dependent variable has its own expectation as an explanatory variable. Let us further recall that expectations appear for four variables: capacity utilization, real in-

of large-scale macroeconometric models, in: C. Hargreaves (ed.), *Macroeconomic Modelling of the Long Run* (E. Elgar, Aldershot, 1992).

terests on the public debt, the real exchange rate and the inflation rate. Owners of the model use fixed-value terminal conditions for capacity utilization and for the inflation rate, the values in question being set by exogenous variables. The terminal condition imposed on each of the other two variables is generated from the paths of "long-run equilibrium" variables which is determined endogenously: the condition is for the variable in question to grow at the same rate in the last period as does the corresponding equilibrium variable[241].

It turns out that the solution path exhibits very little sensitivity to the choice of the horizon, so that a length of eight years, say, is already sufficient for stabilization of the results concerning the first five years. But the adequacy of the terminal conditions in comparative policy analyses may be questioned. In particular the terminal value of the inflation rate is set by an exogenous variable proxying the equilibrium of the ratio "public sector borrowing requirement"/GDP. In evaluating some policy variants in which the exogenous variable in question was properly revised, the adjustments appear to be accurate. The same conclusion applies for simulations which do not imply a change in the long-run inflation rate. But Fisher convincingly argues that things are not always as simple: some simulations and their attendant revisions of the exogenous variable may even alter the stability properties of the model.

4.3.4. *Policy analysis with one eye to model-consistent expectations the other to adaptive expectations*[242]

Large-scale quarterly macroeconometric models have been used since the late 1960s at the US Federal Reserve Board for forecasting and quantitative analysis. Whereas a number of changes had been brought to the initial models throughout the years, the project of a basic redesign was initiated in the early 1990's. In 1996 the earlier generation of models was officially replaced by a new one, differences lying primarily in the specification of expectations and the process of dynamic adjustment to long-run equilibria (in the short run prices are sticky and output is determined by aggregate demand, in the long run prices fully adjust and equilibrium is determined by supply factors).

The main policy model contains now over 250 behavioural equations, of which 40 describe the US economy and the rest describes the 11 other countries and regions. Anticipated values of future variables directly influence prices,

[241] This concerns the "Autumn 1985 version" of the model, the one with which were run the policy simulations reported above. In his book Fisher also discusses the serious problems which arose from different terminal conditions used in an earlier version: the values of the two variables in question had to exactly match in the last period those of the corresponding equilibrium variables.

[242] We are drawing here from F. Brayton, A. Levin, R. Tyron and J. Williams (1997), op. cit.

wages, interest and exchange rates, as well as various components of aggregate demand. "Recognizing that no single assumption regarding the formation of expectations is likely to be appropriate in all circumstances and that it may be useful to see how different specifications in this regard affect system properties, the new model has been designed to have the flexibility to be simulated under alternative assumptions about how expectations are formed. At present, two options (or combinations thereof) are available: backward-looking, or adaptive, expectations and model-consistent, or rational, expectations". For the US the main assumption about the generation of adaptive expectations is that all agents share a common small VAR model of the economy that includes past observations of consumer price inflation, the short-term interest rate, the output gap, as well as long-run expectations of inflation (survey-based) and of interest rates (forward rates). For the other countries, adaptive expectations are extrapolations of univariate processes.

A central specification for dynamic adjustments inspires the choice of some twelve to fourteen US structural equations: for inventories, fixed investment, prices, wages,... Those are the equations containing explicit expectations. For a variable, y_t say, the central specification would contain lagged values $y_{t-\tau}$, an error–correction term involving the value y^*_{t-1} of a desired level for y_{t-1} and expectations $y^{*e}_{t+\tau}$ about this desired level. More precisely the structural equation would write:

$$\Delta y_t = A(L)\Delta y_{t-1} + b\big(y^*_{t-1} - y_{t-1}\big) + C(F)\Delta y^{*e}_t, \tag{294}$$

Δ being of course the first difference operator, b a number, $A(L)$ a polynomial of the lag operator and $C(F)$ a polynomial of the lead operator ($F^m x_t = x_{t+m}$). In such dynamic adjustment equations the polynomial $A(L)$ and $C(F)$ are most often found to be both significant: in other words, lags and leads are both important for adjustment behaviour.

A number of policy simulations were made with this model. For instance, a set of simulations concerned responses of the US economy to a given monetary policy shock raising for one quarter the US short-term interest rate above what the normal policy implied and keeping unchanged the long-run inflation objective of the policy. Graphs for the two simulations, with respectively model-consistent and adaptive expectations, are qualitatively similar: during about a year, a rise in the long-term interest rate and, during about three years, a depression of output and a decrease in the inflation rate, maximum effects occurring after a year. But the price effect is much more important and somewhat more persistent with adaptive than with rational expectations, correlatively the output effect is less important. In particular with such a purely transitory policy, heuristic arguments would have led us to predict the same kinds of differences between the two simulations (see, for a different policy change, our heuristic arguments in Subsection 4.3.4 of Chapter 8).

The authors also aimed at assessing the performance of their model, used with respectively the two hypotheses about expectations, compared with what could be inferred from a statistical analysis of historical data. The latter analysis was precisely borrowed from what was normally done in the RBC movement for the characterization of "stylized facts". Model simulations were computed with shocks randomly drawn from the series of the values taken by the vector of residuals in a fit of the model to the data covering the years 1966 to 1995. Comparison between the simulation results and historical stylized facts bore on the empirical variances of the fluctuations of the main variables, on their correlations with the fluctuations of GDP, and on their lagged correlations with those fluctuations.

Results exhibit a high degree of similarity between the three tables, about the variances, correlations and lagged correlations, for respectively the historical data, the simulations assuming rational expectations and those assuming adaptive expectations. However, rational expectations lead to overpredict the amplitude of GDP fluctuations, whereas adaptive expectations lead to underpredict it. More importantly there is a striking contrast between the results of the expectations assumptions in the block concerning prices and wages. Under adaptive expectations the match between the simulated moments of price inflation and those from historical data is close; the simulated real wage is procyclical as it is historically, but slightly less so. On the other hand, the outcome under model-consistent expectations is far less favourable: the real wage is mildly countercyclical and the correlations across time of the rate of inflation with output are generally inconsistent with the historical correlations.

Summarizing what was learned thus far, in particular from the UK and the Federal Reserve Board experiences, we have to remember that rational-expectations macroeconometric models do not necessarily give more accurate predictions than structural models developed in the adaptive-expectations tradition. But in policy analysis they open possibilities which are not available with those models, namely to distinguish between transitory and permanent policy changes, between anticipated and unanticipated decisions, between perfectly and imperfectly credible policy announcements (Brayton et al. report an interesting analysis of effects of this last distinction).

4.4. The VAR approach to stochastic modelling

Vector autoregressions are now among the most widely used tools in time series macroeconometrics. They retain interest because of the attraction of their deliberately empirical approach and their claim to use only minimal sets of prior assumptions. At the same time recurrent doubts are also expressed about some of their features, hence about their significance and the place they should receive in efforts for the improvement of macroeconomic knowledge. Basically this ambiguous situation reflects the difficulty of using empirical evidence in economics, where controlled experiments are seldom feasible and where a large part of our information about the facts comes from what we directly know of institutions, constraints and behaviours. The ambiguity in question is a natural consequence of the eclecticism which protects economists from too long misinterpreting parts of their non-experimental observations. In any case a textbook on macroeconomics should not ignore the trends along which the VAR technique and its use are developing.

Actually the VARs made their appearance at various places in our Chapters 8 and 9. They were introduced in the second half of Part 5 of Chapter 8, when we discussed the literature on the effectiveness of monetary policy: the elements of the technique were presented at this first occasion (Section 5.5). In Section 7.5 of the same chapter we wondered whether and how we could use VARs in order to best detect the shocks to which our economies were exposed. In Sections 3.4 and 3.5 of Chapter 9 VARs were occasionally taken as providing benchmarks for evaluating the usefulness and properties of structural models. Finally, in Section 3.8, the VARs were naturally associated with the criticism addressed by C. Sims to structural simultaneous-equations models as presented in Part 3 of this chapter.

The reader remembers in particular that, in his key article[243], C. Sims argued that, in order to permit identification of the parameters of these structural models, macroeconomists had to use strong restrictions, which he described as incredible. But we also saw that identification of the parameters was crucial for the application of the VAR techniques and that there was large room for discussing the realism of the restrictions used to this effect by C. Sims and other econometricians. No one should then be surprised to learn that the identification question remains, after twenty years, the main issue, concerning not only the domain to allocate to VAR techniques but also the adaptation to be brought to these techniques in order to best suit the needs of particular applications. This is so much so that today consideration of the so-called "structural VARs" provides the best entry to the study of recent developments concerning the use of VARs in macroeconomics.

The phrase "structural VAR" appears natural when we reflect on the fact that many linear dynamic macroeconomic models have a VAR representation

[243] C. Sims, Macroeconomics and reality, *Econometrica* (January 1980).

and that, in many cases, the properties of a multidimensional process ruled by a non-linear model may also be usefully approximated by those of a VAR. The remark applies not only to structural models as developed in the decades following their introduction by J. Tinbergen and L. Klein, but also to the models worked out by the real-business-cycle movement. But the parameters of dynamic macroeconomic models are most commonly subject to many restrictions, which appeared when the model was built. Some of these restrictions at least may be used in order to identify the parameters of the VAR representation. Many macroeconometricians were attracted by such an idea when aiming at finding in the data the VAR representation that would empirically give the respective strengths of various lags in the relations between economic variables.

There is some leeway in the transfer of restrictions from a model to a VAR, either in the choice of which restrictions to select for that, or even in the interpretation to give to a restriction when it is moved from one context to the other. Hence, some econometricians prefer not to speak of structural VAR in the description of what they do, but rather of "semi-structural VAR" or of "model-based VAR-methodology". We easily understand that individual practices with respect to such a transfer vary, depending on the answers given to questions such as: how confident should we be in the realism of the economic foundations underlying the model? Is it sufficient for the econometrician to deliver a model that accords well to his or her data?

As soon as we deviate from the original stand taken by proponents of VAR analysis, then seen as an atheoretical way of estimating the dynamic impacts of economic shocks, as soon as we use prior economic ideas to identify a VAR, the distinction between the VAR modelling strategy and the standard simultaneous equation approach becomes a matter of degree, not of principle. The issue then centers on which restrictions can be viewed as most credible (or least incredible, according to tastes).

4.4.1. *Introduction to structural VARs*

The idea to find the identifying restriction of a VAR analysis in a structural model is usually traced back to two articles published in 1986, one by O. Blanchard and M. Watson[244], the other by B. Bernanke[245]. Let us look at these two articles, which well demonstrate the potential of the idea and its limitations.

[244] O. Blanchard and M. Watson, Are business cycles all alike?, in: R. Gordon (ed.), *The American Business Cycle* (University of Chicago Press, 1986).

[245] B. Bernanke, Alternative explorations of the money-income correlation, *Carnegie Rochester Conference Series in Public Policy* **25** (1986) 49–100.

(i) In Section 7.5 of Chapter 8 we saw which interesting conclusions Blanchard and Watson derived from their four-dimensional VAR analysis about the characteristics of the shocks to which the US economy had been exposed between 1947 and 1982. We then postponed, however, until the present section study of the identifying restrictions they used. These restrictions are derived from the following small static structural model linking the four variables of the analysis: y the logarithm of real GNP, p the logarithm of the GNP deflator, m the logarithm of nominal M_1, g an index of the effect of fiscal policy on aggregate demand; the equations then are:

$$y_t = a_{12}p_t + a_{13}m_t + a_{14}g_t + \eta_{1t}, \tag{295}$$

$$p_t = a_{21}y_t + \eta_{2t}, \tag{296}$$

$$m_t = a_{31}y_t + a_{32}p_t + \eta_{3t}, \tag{297}$$

$$g_t = a_{41}y_t + a_{42}p_t + \eta_{4t}. \tag{298}$$

The four structural Equations (295) to (298) are meant to be respectively the aggregate demand equation, the supply equation, the monetary policy rule and the fiscal policy rule. The unobservable vectors η_t with components η_{1t} to η_{4t} are supposed to be serially uncorrelated and to have a time-invariant diagonal covariance matrix D.

Let us note in passing that the authors are rather brief about the restrictions contained in this structural model. About the diagonality of D they write: "Although we do not believe this is exactly the case, we find it plausible that [relevant disturbances] have a low correlation, so that our identification is approximately correct". They further write: "We have chosen standard specifications for aggregate supply and demand... The policy rules [allow contemporaneous responses] to output and the price level".

In order to see how the restrictions contained in the model (295) to (298) are transposed into identifying restrictions of a vector autoregression, let us go back to the notation used in Section 5.5 of Chapter 8, where the equation then numbered as (223) applied to the multidimensional process of the observed vector y_t. The equation was:

$$\sum_{\tau=0}^{h} A_\tau y_{t-\tau} = A(L)y_t = \varepsilon_t. \tag{299}$$

(We keep here the notation y_t for the vector appearing in general specifications such as the one above, since there is no risk of confusion with y_t designating output in the display of the small structural model.) Blanchard and Watson derive from their model restrictions on the matrix A_0 appearing in (299) above:

for instance, in the second line $a_{0,23} = a_{0,24} = 0$. The diagonality of D, joined to the fact that one of the coefficients in each of the four equations is normalized to be equal to 1, is equivalent to the assumption made in Chapter 8 and stating that the covariance matrix of the vector ε_t was equal to the identity matrix.

At this point we see that we are close to the recursiveness hypothesis often made in VAR analyses (see Sections 5.5 and 5.6 of Chapter 8), except that the zero restrictions imposed here on the matrix A_0 do not make this matrix triangular. The matter is even a little more complex, because triangularity of a (4×4) matrix imposes that 6 elements are equal to zero, whereas in (295)–(298) there are only 4 zero coefficients (a_{23}, a_{24}, a_{34}, a_{43}). Identification requires at least two extra restrictions. Blanchard and Watson find them in imposing the values of a_{41} and a_{42} in the fiscal policy rule. They write: "Within a quarter, there is little or no discretionary response of fiscal policy to changes in prices and output. Most of the response depends on institutional arrangements, such as the structure of income tax rates, the degree and timing of indexation of transfer payments, and so on. Thus the coefficients [a_{41} and a_{42}] can be constructed directly". The authors do so in an appendix of their article and take the values then obtained for the two coefficients as fixed in their VAR analysis, the results of which were presented in Section 7.5 of our Chapter 8.

In his comments published after the article, R. Shiller is quite critical of the identifying restrictions used by Blanchard and Watson. He writes in particular: "where did the assumption [that the error terms are uncorrelated across equations] come from? One way of appreciating the arbitrariness of such an assumption is to note that if the model holds for monthly data, say, with error terms uncorrelated across equations, then the quarterly data will not generally have a representation with error terms uncorrelated across equations". He also writes: "Another identifying assumption is that monetary policy does not depend on the contemporary fiscal policy variable. Why is this omitted? Does the Fed always ignore fiscal policy?". However, in the discussion C. Sims replies to R. Shiller: "Such a criticism does not mean that identification should not be attempted; it means only that identifying restrictions should be used wisely, as Blanchard and Watson have done".

(ii) The article by B. Bernanke (op. cit.) aimed at testing whether the empirical correlation between the money stock and national income could not be better explained by the "credit view" (see Section 2.10 of this chapter) than by the "real business cycle view", which takes the position that money is passive and that it is correlated with output only because agents increase their transaction demand for money when output is high.

The methodology is quite similar to that of the article by Blanchard and Watson, to which Bernanke refers. There is, however, a significant difference, which we may best see on the general specification. Instead of writing (299) and assuming that the disturbance vector ε_t has a diagonal matrix, Bernanke keeps the same orthogonality assumption but applies it to the disturbance vector u_t which enters the system written as:

$$A(L)y_t = \Gamma u_t \tag{300}$$

where the square matrix Γ is no longer the identity. This matrix is subject to restrictions given by the structural model, which also gives the restrictions on A_0. In order to explain this difference, Bernanke writes: "I think of the u_{it} as primitive exogenous forces, not directly observed by the econometrician, which buffet the system and cause oscillations. Because these shocks are primitive, i.e., they do not have common causes, it is natural to treat them as approximately uncorrelated. However, one would not want to restrict individual u_{it} to entering one and only one structural equation in general".

In the application discussed the vector y_t of observed variables has six components, all of them in logarithms. Three of these variables are the same as in the Blanchard–Watson application, the fourth g_t being real defense spending, the fifth b_t the monetary base, the last c_t the sum of credits granted by banking institutions. The following static structural model is posed:

$$y_t = a_{13}(m_t - p_t) + a_{14}g_t + a_{16}(c_t - p_t) + \gamma_{12}u_{2t} + u_{1t}, \tag{301}$$

$$p_t = a_{21}y_t + u_{2t}, \tag{302}$$

$$m_t = a_{31}y_t + a_{32}p_t + a_{35}b_t + u_{3t}, \tag{303}$$

$$g_t = u_{4t}, \tag{304}$$

$$b_t = a_{51}y_t + a_{52}p_t + a_{53}m_t + a_{54}g_t + u_{5t}, \tag{305}$$

$$c_t = a_{61}y_t + a_{62}p_t + a_{65}b_t + u_{6t}. \tag{306}$$

We shall not repeat here the arguments given for the zero restrictions imposed on elements of the matrix A_0, but simply note that the matrix Γ is assumed to have a single non-zero off-diagonal element: γ_{12}. In other words, all variables except output would be each subject to just one primitive specific shock, but the price shock (alternatively "supply shock") could add an impact on output to the impact of the specific demand shock. The argument then given by the author runs as follows.

He has three reasons to choose a just-identified structural model: first, this is convenient for estimation; second, overidentified models will not in general yield orthogonal residuals, a fact that would clash in interpretations of the results with the assumed orthogonality of the u_{it}; third, the use of just-identified models tends to minimize the number of "auxiliary" restrictions, which "are not strongly implied by the basic theory". This choice being made, the author observes that, in the case studied, there remains just one degree of underidentification after specification of the restrictions on A_0. Therefore Γ should have just one non-zero off-diagonal element; the choice of γ_{12} as being this element is advisable because "it allows non-zero correlation between the disturbances to the model's two principal behavioral equations". Once again we realize that the choice of identifying restrictions is delicate, and not necessarily persuasive for others than the author of a VAR analysis.

Before proceeding to the main analysis estimating the VAR model (300) with the restrictions on A_0 and Γ implied by the structural model (301)–(306), Bernanke fitted on his data base a standard VAR using the common recursiveness assumption. He could so compare his preferred estimates to that standard. He was so led to assert: "It appears that credit shocks are important for output, the inability of the standard VAR methodology to find this being due to its failure to separate correctly the" truly exogenous "component of credit [u_{6t}] from its endogenous part... The new estimates do not imply that the credit channel has replaced the monetary channel; instead, money and credit are parallel forces of approximately equal importance". The conclusions of the article are cautious: "For policy makers, the conservative course is to continue to use care to avoid destabilizing shocks to credit markets and to the money supply... The results of this paper should be considered as tentative. The paper does demonstrate, however, that the structural interpretation of VARs can be very sensitive to the model that one assumes (implicitly or explicitly) is relating contemporaneous residuals".

4.4.2. *About the choice of identifying restrictions*

Having realized the difficulty and practical importance of the choice of identifying restrictions in VAR analyses, let us now briefly stop and collect our ideas about it. In Section 5.6 of Chapter 8 we discussed for the first time macroeconomic applications of vector autoregressions, then intended to empirically evaluate the effectiveness of monetary policy. The applications we saw in that section all used the recursiveness identifying assumption (A_0 triangular, D diagonal), but not all with the same ordering of the variables: whereas the results drawn from the study of E. Leeper, C. Sims and T. Zha (op. cit.) used the ordering placing first the indicators of monetary policy (M, r), the order selected by L. Christiano, M. Eichenbaum and C. Evans (op. cit.) placed first three macroeconomic variables concerning the activity of the private sector (in particular y and p). We noted then, however, that the two studies led to roughly similar estimations of impulse-response functions, at least over a period of four years after the shocks. In Section 7.5 of Chapter 8 we examined the results of VAR analyses aimed at characterizing the shocks to which modern economies are exposed. The first study there, due to O. Blanchard and M. Watson, used as identifying restrictions orthogonality of the disturbances in (299) and the zero restrictions on elements of the matrix A_0 which followed from the structural model (295)–(298). We just saw how Bernanke relaxed the orthogonality assumption when replacing (299) by (300), but soon after introduced in his application the strong restrictions on the matrix Γ which were suggested by his model (301)–(306). To complete our survey of the identifying restrictions considered in this book up to this point, we must still look at the analysis by O. Blanchard and D. Quah (op. cit.), the second of those studied in Section 7.5 of Chapter 8.

A common feature of all the applications listed above was to leave unrestricted all matrices A_τ in (300) except A_0. (When we say this we ignore, however, the restrictions implied by the limitation to a selected set of variables for simultaneous analysis, as well as by the limitation to the number h of lags appearing in the autoregression.) Nothing in principle prevents introducing restrictions on other matrices A_τ if we have good reasons for making assumptions of this sort. We may even consider restrictions involving jointly several matrices A_τ.

Blanchard and Quah indeed made an assumption on the whole series of A_τ through a restriction expressed on the matrices C_τ appearing in the moving-average representation of the autoregression (the representation was Equation (224), corresponding to the direct form (223) of the autoregression, both equations being in Chapter 8). We remember that the restriction concerned the *long-run* response of output to shocks meant to hit aggregate demand rather than aggregate supply (the reasons for the exact form of the restrictions were explained in Section 7.5 of Chapter 8; they will not be repeated here).

We remember that some reservations could be found against the hypotheses chosen by the authors. Here, let us simply recall those concerning the idea of relying on assumptions about what the long-run responses ought to be, an idea which occurred in other VAR analyses. Two remarks may be troublesome. First, long-run responses are actually determined by the whole dynamics of the process studied; they may then depend on phenomena involving long lags, such as the replacement of equipments to be scrapped in the future, or the productivity gains resulting from the experience progressively acquired on the job by the labour force. It is hazardous to assume that the long-run effect of all such long-lag causalities is either globally negligible or well captured by an initial autoregression involving lags covering only a few quarters. Second, VAR analyses actually bear on detrended series; it is not very likely that the choice of the method used for computing and removing the trend has no effect on the subsequent estimation of long-run responses[246]. These two remarks may explain why the identifying restrictions most commonly used are specified on the matrix A_0 only, or on the two matrices A_0 and Γ of Equation (300).

Still another interesting possibility for identification is to refer to the reduced form following from the latter equation, namely:

$$y_t = -\sum_{\tau=1}^{h} A_0^{-1} A_\tau y_{t-\tau} + A_0^{-1} \Gamma u_t. \tag{307}$$

The initial autoregression fitting equation (225) of Chapter 8 gives estimates of the matrices $-A_0^{-1}A_\tau$ (the matrices B_τ^* in that equation) and residuals. Knowing these residuals, possibly accepting orthogonality of the elements of u_t, we can estimate the matrix $A_0^{-1}\Gamma$ if we have enough outside information about this matrix. Using recursively Equation (307) written with the estimated matrices, we can then compute estimates of the response of each $y_{j,t+\theta}$ to each supposed unit shock u_{it}. The estimated impulse-response function of y_j to shocks on u_i is given by the series so obtained for all non-negative values of θ. The outside information may conceivably be found in "economic theory", but it may also come from other sources of prior information about the phenomenon. We shall presently see an application which is illustrative in this respect.

4.4.3. *Contributions to macroeconomic knowledge: impact of fiscal policy*

Providing a survey of the many uses of VARs in macroeconomics would go beyond the knowledge of this author. In order to give an idea of the variety of topics that would appear in such a survey, we must, however, try to selectively span the wide range of applications which can be found in the literature. We saw in Chapter 8, Sections 5.6 and 5.7, how VARs contributed to the study of the effects of monetary policy, and we shall still see more about the same subject in Subsection 4.4.5. We saw in Chapter 8, Section 7.5, how VARs were addressed to the quite relevant question of knowing to which shocks are modern economies exposed. Now we shall complement the picture by considering an application of VARs to the study of the effects of fiscal policy. In the next subsection we shall look at another application exhibiting a more multivariate study of macroeconomic dynamics, in the spirit of business-cycle research.

[246] For a fuller discussion of these difficulties see Section 4.2 in F. Canova, Vector autoregressive models: specification, estimation, inference and forecasting, in: H. Pesaran and M. Wickens (eds.), *Handbook of Econometrics: Macroeconomics* (Blackwell, Oxford, 1995).

O. Blanchard and R. Perotti[247], coming after a few other macroeconomists who had already used VARs for the study of the macroeconomic impact of the US fiscal policy in the postwar period, wanted to conduct a somewhat more comprehensive exploration of the same topic. Their paper is particularly interesting for us at this point because their approach "relies on institutional [outside] information about the tax and transfer systems and the timing of tax collections to identify the automatic responses of taxes and spending to activity and, by implication, to infer fiscal shocks".

They mainly study the simultaneous quarterly variations in three variables: GDP, government spending (only in purchases of goods and services) and net taxes (taxe revenues *minus* transfers), all three variables in real per capita terms and in logarithms. Let now the vector ξ_t be the error term in the reduced-form autoregression model, the one to which the initial fit applies, in other words the last term in (307). The three elements of this vector are denoted as x_t, g_t, t_t; they are called by the authors "unexpected movements" in respectively output, spending and taxes. Correspondingly the three primitive shocks, called by the authors "structural shocks", may be denoted here: u_{xt}, u_{gt}, u_{tt}. Blanchard and Perotti further pose:

$$x_t = a_{12}g_t + a_{13}t_t + u_{xt}, \tag{308}$$

$$g_t = a_{21}x_t + u_{gt} + b_{23}u_{tt}, \tag{309}$$

$$t_t = a_{31}x_t + u_{tt} + b_{32}u_{gt}. \tag{310}$$

The three shocks u_{xt}, u_{gt}, u_{tt} are assumed to be contemporaneously uncorrelated, for the same reasons as those given by Bernanke in the article discussed within Subsection 4.4.1 here.

Before looking at the restrictions embodied in System (308)–(310) let us connect the approach to what we discussed earlier. The system is relating the vector ξ_t of errors in the reduced-form autoregression, i.e. the vector of unexpected movements, to the vector u_t of structural shocks. In Equation (307) the relation is made explicit with:

$$\xi_t = A_0^{-1}\Gamma u_t. \tag{311}$$

We may write it:

$$A_0\xi_t = \Gamma u_t \tag{312}$$

an equation which may also be considered as directly derived from the structural-form autoregression (300). But knowing the nine elements of the

[247] O. Blanchard and R. Perotti, An empirical characterization of the dynamic effects of changes in government spending and taxes on output, NBER working paper series No. 7269 (July 1999).

matrix $A_0^{-1}\Gamma$, which is the only requirement for computing the structural shock-response functions from the estimation of the reduced-form autoregression, does not need full knowledge of the eighteen elements of A_0 and Γ. Reference to the structural form may be convenient for inserting identifying outside information; but it does not seem to be absolutely necessary[248]. Let us keep the remark in mind and move ahead in our examination.

Reading Blanchard and Perotti we clearly see that they mean their system (308)–(309) to correspond to Equation (312) with its structural interpretation. We may then write:

$$A_0 = \begin{bmatrix} 1 & -a_{12} & -a_{13} \\ -a_{21} & 1 & 0 \\ -a_{31} & 0 & 1 \end{bmatrix},$$

$$\Gamma = \begin{bmatrix} 1 & 0 & 0 & \\ 0 & 1 & 0 & b_{32} \\ 0 & b_{32} & 1 & \end{bmatrix}. \tag{313}$$

The two diagonals are pure normalizations, for A_0, in the writing of structural equation and, for Γ, in the convention that the unit of measurement of any structural shock is the same as the unit of measurement of the corresponding observed variable. The two zeros imposed in the matrix A_0 may similarly appear natural for the expression of the fiscal policy rule, which determines jointly spending and taxes.

Their knowledge of US institutions leads the authors to write: “Direct evidence on the conduct of fiscal policy suggests that it takes policy makers and legislatures more than a quarter to learn about a GDP shock, decide what fiscal measures to take in response, pass these measures through the legislature, and actually implement them”. This remark explains the two zeros in the first column of Γ: the policy shocks ignore the contemporaneous activity shock u_{xt} because no discretionary adjustment is made to fiscal policy in response to unexpected events within the quarter. However, the terms $a_{21}x_t$ and $a_{31}x_t$ capture the automatic effects of activity on spending and taxes under existing fiscal codes and regulations. It is then possible to look in details at the institutional dispositions and to infer from them direct measures of the coefficients

[248] We may note in passing a result which does not apply with the present model but may have interest in other applications. If we knew fully the matrix $\Gamma D \Gamma'$, for instance, that it is the identity matrix, and if the restrictions on A_0 were all linear and implying a particular form of block-recursiveness, the impulse response functions to policy shocks, as computed through (313) from an estimated covariance matrix Σ would be independent of the choice of A_0 among the matrices fulfilling the restrictions. This is result (ii) of Proposition 4.1 in: L. Christiano, M. Eichenbaum and C. Evans, Monetary policy shocks: what have we learned and to what end?, NBER working paper, No. 6400 (February 1998).

a_{21} and a_{31}. The result is introduced as outside information in the matrix A_0. We may so say that, within the VAR analysis, the matrix A_0 depends on just two unknown parameters a_{12} and a_{13}.

We then turn attention to Γ where we note that the model assumes two zeros in the first line: Equation (308) says that unexpected movements in output do not depend on structural policy shocks: this is easily explained again by lags: in the definition and implementation of policies, as well as in the definition and implementation of private decisions. So, matrix Γ has two unknown elements b_{23} and b_{32}.

But, at this point the authors realize that they have not yet enough identifying restrictions. Indeed, in order to infer estimations of A_0 and Γ from (312) they are going to look at the estimation of the covariance matrix Σ of the unexpected-movement vector ξ_t, taking moreover account of the assumed orthogonality of the structural-shock vector u_t, i.e. of the diagonality of its covariance matrix D. The identifying equations will then be:

$$A_0 \Sigma A_0' = \Gamma D \Gamma'. \tag{314}$$

Equality of the two symmetrical matrices of order 3 entails 6 independent equations. But these equations contain 7 unknowns: the 3 diagonal elements of D, together with a_{12}, a_{13}, b_{23}, b_{32}. The authors then decide to make two "agnostic" computations, the first with the additional restriction $b_{23} = 0$, the second with $b_{32} = 0$. It turns out in the application that, "in nearly all cases", the correlation between u_{gt} and u_{tt} is sufficiently small for the two computations to give similar impulse responses of output, both of which are given in the article.

The main application, the only one to be reported here, uses the US quarterly time series for the years 1960 to 1997. The accuracy of the results is unfortunately weak, a fact leading the authors to present confidence bands for just one standard error on each side of the estimated response functions (this low degree of aspiration about what can be called significant seems to be now frequent in applications of the VAR approach). The positive government spending shocks are found to have a positive effect on output during the first year, with a multiplier of the order of 0.7; the same order of magnitude applies in subsequent quarters, but with such a low accuracy as to be hardly worth remembering. Except for the opposite signs, tax increases being depressive, about the same summary can be given of the estimated effects of tax shocks. The authors also present results of four-variable VARs, in each one of which a particular component of aggregate demand is added in the analysis, with the same restrictions as those appearing in (308). A surprising result is noted: government spending shocks are found to induce a persistent decrease in private investment spending.

Did this analysis add much to our knowledge of the impact of fiscal policy? Probably not, because its results turn out to be so imprecise. Perhaps this could be explained by macroeconomists who learned from structural macroeconometric models, which embody more information than is contained in aggregate time series. When we discussed in Section 3.5 our knowledge about the values of multipliers, particularly at the end of Subsection 3.5.1, we saw that these values much depended not only on disequilibria in the situation when fiscal policy intervened, but also on how this policy was financed. For instance, results obtained by R. Fair predicted that in the US the short-term multiplier after a permanent increase in government spending would be equal to 1.2, 3.2 or 1.6 depending on whether it was financed entirely by the issue of bonds, or entirely by the issue of money, or by both in such a way that the interest rate remained unchanged. This may suggest that, to a considerable extent, the inaccuracy in the results here discussed may be due not only to the inherent linearity of VAR analyses, hence to their neglect of changes in disequilibria, but also to the fact that the analysis treated fiscal policy in isolation, independently of monetary policy.

4.4.4. *Contributions to macroeconomic knowledge: stylized facts*

When in Section 5.6 of Chapter 8 we examined applications of the VAR approach to the diagnosis about the effectiveness of monetary policy, we deliberately looked only at applications involving very few macroeconomic variables. We mentioned, however, that within the same projects impulse-response functions were also estimated for larger sets of variables and that those more extensive results were interesting for business cycle theory. We must now look at some such results. For the purpose we take again the article of L. Christiano, M. Eichenbaum and C. Evans[249].

The authors use large data bases of quarterly US time series, inserting in turn each one of the series they select into a reference core autoregression. More precisely, the latter autoregression contains six variables: three macroeconomic variables (real GDP, the GDP deflator, an index of sensitive commodity prices) and three monetary variables (the "federal funds rate of interest", controlled by the Fed acting as central agent of day-to-day borrowings and lendings between commercial banks, the regulated "total reserves" of commercial banks at the central bank, the "non-borrowed reserves", i.e. the part of

[249] L. Christiano, M. Eichenbaum and C. Evans, The effects of monetary policy shocks: evidence from the flow of funds, *Review of Economics and Statistics* (February 1996).

total reserves which is not covered by commercial banks borrowings from the Fed for short periods at the "discount window"). (As we saw in Section 5.6 of Chapter 8, the impulse-response functions are identified by the recursiveness hypothesis.)

About any other series, for instance, employment, the authors compute a VAR with seven variables, employment being added last in the recursive order after the six variables of the core autoregression. They publish results about the response of employment to shocks on the main monetary policy instrument (alternatively the federal funds rate, or the non-borrowed reserves). Incidentally we note that, except for the identifying restrictions, this is precisely the procedure used by Blanchard and Perotti when they inserted in turn into their autoregression the main components of aggregate demand.

Let us reflect on the nature of the so computed responses of employment to monetary policy shocks. Readers looking at the graph of the impulse-response function will certainly tend to interpret them as "stylized facts": after a discretionary transitory one percentage-point increase in the annual rate of interest on federal funds, and with a lag of about a semester, employment will decrease up to -0.4% at the end of the second year, before it starts slowly increasing again (the initial shock actual means that the rate of interest will still be higher by 0.75 point after six months, by 0.25 point after a year). Similarly, the reader will note than such discretianory transitory increases account for 17% of the variance of the unavoidable error that would be made in a forecast of employment six years ahead. The rest of the variance is accounted for, not only by shocks hitting directly employment, but also by the shocks on the five other variables of the analysis, on output in particular.

How do such results compare with, or complement, those delivered by the "new methodology" introduced in the 1980s for the systematic empirical characterization of "business cycle facts" (see Section 2.2 in this chapter)? Essentially in the same way as multivariate correlation analysis compares with, or complement, bivariate correlations. The stylized facts exhibited by Clarida, Eichenbaum and Evans are the result of a deeper examination of simultaneous covariations than those shown by Backus–Kehoe, Danthine–Donaldson, Fiorito–Kollintzas and others. Their descriptive value is undisputable, even though these stylized facts stop short of providing a fully multivariate analysis of the large data bases from which they come. The authors indeed explain that, with their series of 7-variables autoregressions, they chose an "intermediate strategy" because estimating autoregressions with larger numbers of variables soon become impossible; but they recognize that their results are exposed to omitted variable bias. The explanatory value of these results is moreover subject to other reservations applying generally to inferences drawn from VAR analyses.

This being said, let us take a brief overview of the results, speaking loosely in terms of causality and limiting attention to response fonctions to impulses on federal funds rate[250].

We first note that, after a delay of about two quarters, a contractionary policy shock leads to a manifest decrease in output, to a somewhat weaker decrease in employment and to a still weaker increase in unemployment. However, the response is immediate in some variables such as the fall in retail sales and temporary build-up in manufacturing inventories during four quarters. After a semester again, the real wage decline appears and seems to be persistent; it is manifest in the second year, particularly in manufacturing durable goods industry. Profits after taxes fairly quickly decrease, but recover in the third and fourth years.

For business cycle theorists who stress the importance of credits it is interesting to look at the results concerning flows of funds. Movements in households funds are not strongly characterized, except to say that they seem slow to react to monetary policy shocks. In contrast the reactions of the business sector to a contractionary policy shock are manifest in the movements of both assets and liabilities, particularly in corporations and in large manufacturing firms. Real net funds raised by enterprises increase during the first three quarters, but decrease during the second and third years, this being due more to changes in borrowings than in lendings, and concerning almost exclusively short-term funds.

4.4.5. *Contributions to macroeconomic knowledge: monetary policy*

The use of vector autoregressions for the study of the effects of monetary policy was the main focus of attention in Sections 5.6 and 5.7 of Chapter 8. We also saw, in Subsection 4.4.1 above how one of the two articles which introduced structural VARs in 1986, the one by B. Bernanke, was dealing with the impact of monetary policy. During the last years a number of research projects went on with the same purpose. We might argue that they are moving more and more into the domain of "Money and Banking", which stands outside macroeconomics narrowly understood (as it was when the scope of this book was announced to exclude financial economics). But, in order to give a proper balance to this section, we cannot overlook this active field in which VARs seem to be appropriate at the border between macroeconomics and banking. In this spirit we shall still look at an interesting article, among others.

[250] Some of the results are found in L. Christiano, M. Eichenbaum and C. Evans, Sticky price and limited participation models of money: comparison, *European Economic Review* (June 1997).

B. Bernanke and I. Mihov[251] aim to measure the monetary policy stance by VAR estimation techniques applied in conjunction with prior information about central bank operating procedures. They are motivated by the observation that various measures are used, such as the federal funds rate, or non-borrowed reserves, or still total reserves in L. Christiano, M. Eichenbaum and C. Evans (1996, op. cit.). The two authors say that there is little agreement on which measure most accurately captures the stance of policy, and that the choice matters for a quantitative assessment of the effects of policy. Using simultaneously prior information on operating procedures and time series analysis should permit to resolve the indeterminacy.

Before engaging in their study they think it advisable to remove a worry that may have come to minds of readers in the foregoing subsection when stylized facts obtained from VARs were compared with those obtained in Section 2.2 by direct application of correlation analysis. They write: "A frequently heard criticism of the VAR-based approach is that it focusses on monetary policy innovations rather than on the arguably more important systematic or endogenous component of policy. We believe this criticism to be misplaced. The emphasis of the VAR-based approach on policy innovations arises not because shocks to policy are intrinsically important, but because tracing the dynamic response of the economy to a policy innovation provides a means of observing the effects of policy changes under minimal identifying assumptions. However,... we recognize that it would be useful to have an indicator of the overall stance of monetary policy, including the endogenous component". (The indicator, proposed by the authors at the end of their article, will not be examined here).

Bernanke and Mihov distinguish between variables that are policy indicators and other variables. Let the vector of the first ones be p_t and the vector of the second be x_t. We order the variables in such a way that the correspondence with system (300) holds with in particular:

$$y_t = \begin{bmatrix} x_t \\ p_t \end{bmatrix}, \qquad A_\tau = \begin{bmatrix} B_\tau & C_\tau \\ D_\tau & G_\tau \end{bmatrix}. \tag{315}$$

We may characterize the variables of p_t as being the candidates for providing the policy measures. Actually, the authors consider three such variables, the same ones as were selected by Christiano, Eichenbaum and Evans: the federal funds rate, the non-borrowed reserves and the total reserves.

[251] B. Bernanke and I. Mihov, Measuring monetary policy, *Quarterly Journal of Economics* (1998) 869–902.

The authors introduce in system (300) so partitioned the block-recursiveness assumption defined by:

$$A_0 = \begin{bmatrix} B_0 & 0 \\ D_0 & G_0 \end{bmatrix}, \qquad \Gamma = \begin{bmatrix} \Gamma_x & 0 \\ 0 & \Gamma_p \end{bmatrix}, \tag{316}$$

where the 0 designate zero matrices. The assumption means that neither the structural policy shocks nor the contemporaneous values of the policy variables affect the macroeconomic variables of x_t within the current period (in their application they use monthly data). Moreover the macroeconomic structural shocks affect the policy indicators only through their impact on the values of the contemporaneous macroeconomic variables. These rather innocuous assumptions are important because they permit to estimate the responses of the macroeconomic variables to exogenous policy shocks without having to identify the entire model structure[252].

Identifying G_0 and Γ_p is sufficient, and this can be done by consideration of an equation similar to (312) relating the error terms in the reduced-form autoregression model to the structural shocks. More precisely it is now:

$$G_0 \pi_t = \Gamma_t v_t, \tag{317}$$

where π_t and v_t are the vectors of policy components of respectively ξ_t and u_t (these components are called by the authors "innovations" and "disturbances". In other words, a small static structural model, concerning only the contemporaneous values of the reduced-form autoregression errors (innovations) and the structural shocks (disturbances) of the policy variables will suffice.

In their application the authors build this model from their institutional knowledge of the market for reserves of commercial banks at the Fed. The model contains parameters, but those can be estimated as long as the model is identified by the diagonality of the covariance matrix of v_t and by restrictions on the elements of G_0 and Γ_p. We are going to look at this model, written with notations close to those used by the authors, without pretending to have a sufficient institutional knowledge of money and banking to be sure to have here the right model[253].

We drop the subscript t in π_t and v_t. We write the three components of π_t as π_{FFR}, π_{NBR} and π_{TR} (for innovations in respectively "federal funds rate", "non-borrowed reserves" and "total reserves"). These are linked by three structural equations: a demand equation for total reserves where a structural shock

[252] The mathematical property behind this fact is essentially the same as the one mentioned in Subsection 4.4.3 above and proved in Christiano, Eichenbaum and Evans (1998).

[253] L. Christiano, M. Eichenbaum and C. Evans (1998, op. cit.) write that this model "does not look persuasive neither theoretically nor empirically".

v^d appears, a demand for borrowed reserves (TR-NBR) by commercial banks with a structural shock v^b, a supply of non-borrowed reserves by the central bank (the Fed) with a structural policy shock v^s. More precisely the model is written as:

$$\pi_{TR} = -\alpha\pi_{FFR} + v^d, \tag{318}$$

$$\pi_{TR} - \pi_{NBR} = \beta(\pi_{FFR} - \rho_{DISC}) + v^b, \tag{319}$$

$$\pi_{NBR} = \varphi^d v^d + \varphi^b v^b + v^s, \tag{320}$$

where four parameters α, β, φ^d, φ^b appear as well as the rate ρ_{DISC} at which the Fed lends to banks "at the discount window". Neglecting the presence of the term with this discount rate, this is a particular case of (317) with:

$$G_0 = \begin{bmatrix} \alpha & 0 & 1 \\ -\beta & -1 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \Gamma_p = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & \\ \varphi^d & \varphi^b & 1 \end{bmatrix}. \tag{321}$$

Estimation by an equation such as (314) would cope with the presence of at most 6 independent parameters. But here the equation in question contains 7 parameters, including the three diagonal element of D. At least one more restriction is required. Let us see how the authors justify the three equations (318)–(320) and how they choose the extra restrictions so as to capture various earlier attempts to measure policy shocks.

Equation (318) takes into account a possible dependence of the demand for reserves on the rate of interest that would be earned by money withdrawn from reserves (the price of reserves). In Equation (319) the demand for borrowed reserves "is taken to depend positively on the... federal funds rate (the rate at which borrowed reserves can be relent) and negatively on the discount rate (the cost of borrowed reserves)". Equation (320), which describes the behaviour of the Federal Reserves, assumes that the Fed observes and responds to shocks to the total demand for reserves and to the demand for borrowed reserves within the period, with the strength of the response given by the coefficients φ^d and φ^b. "That the Fed observes reserve demand shocks within the period is reasonable, since it monitors reserves and borrowings continuously... The disturbance term v^s is the shock to policy that we are interested in identifying".

A policy regime, or equivalently a well-specified operating procedure, is characterized by the values of φ^d and φ^b. Particular attention must be given to two cases. If the Fed strictly controls non-borrowed reserves, the observed innovations in the value of these reserves are the policy shocks, which means $\varphi^d = \varphi^b = 0$: we then have the "NBR-model". In contrast, if the Fed strictly targets the federal funds rate its policy shocks must be reflected in π_{FFR}, without interference of the values of the demand shocks. But, adding Equations (319) and (300), and subtracting Equation (318) leads to:

$$\left(\varphi^d - 1\right)v^d + \left(\varphi^b + 1\right)v^b + v^s = -(\alpha + \beta)\pi_{FFR} + \beta\rho_{DISC}. \quad (322)$$

Hence, targeting the federal funds rate requires $\varphi^d = 1$, $\varphi^b = -1$ (the discount rate is infrequently changed). So we obtain the "FFR-model".

A simply way of seeing whether the policy regime is close to a control of non-borrowed reserves or to a federal-funds-rate targeting is to look at the estimates found for φ^d and φ^b when system (317) is identified by $\alpha = 0$ (the demand for total reserves is inelastic in the short run). This particular restriction can be supported by plausible institutional arguments and is convenient. The authors speak of the "just identified model".

We shall not look here at the detailed results obtained by the authors from data for the period 1965–96 and for subperiods. A brief summary runs as follows. The Fed's procedures appear to have changed over time. The FFR model is found to do well for the pre-1979 period and exceptionally well for the post-1988 period; it also appears to be (marginally) the best choice for the whole sample period. The NBR model does well for the "Volcker period" 1979–82, but is otherwise strongly rejected.

The authors also estimate the macroeconomic responses to monetary policy shocks for five distinct models of the market for bank reserves, the three mentioned above and two others also found in the literature. "Qualitatively the results from all five identifications are reasonable, in the sense of conforming to… conventional wisdom": an expansionary monetary policy shock increases output relatively rapidly and raises the price level more slowly but more persistently. However, quantitatively the results differ noticeably. For example, at four years the response of prices under the NBR identification is found to be four times greater than under the FFR identification. These results "underscore the need to choose a model of the Fed's operating procedure that is as nearly correct as possible".

Reflecting on the Bernanke–Mihov's and other recent contributions of VAR analyses to the study of monetary policy, we note a definite selectivity in the choice of those informations about the economic and financial structures that the VAR incorporates. Within the full model, as defined by Equation (300), for instance, the informations so used concern only those elements directly relating to the behaviour of the central bank; they ignore all what may otherwise be known about the private economy. The VAR econometricians may argue that they indeed need not introducing more information because of the mathematical property mentioned here after the hypothesis (316) was made.

But the argument may be misleading. It simply means that no more is necessary, as long as (316) holds and the extra restrictions on A_0 are linear, in order to identify to response to policy shocks from the reduced-form autoregression. In fact, more is necessary if the full model must be estimated, as is the case when the analysis must explain the transmission mechanism thanks to which the responses are obtained. The mathematical property does not hold if some information implies significant non-linear restrictions on A_0 (which would limit the set of admissible impulse-response functions). Finally, there might exist more accurate methods of estimating the responses than those relying on an initial unconstrained reduced-form autoregression. Given these remarks, what was said towards the end of Section 5.7 of Chapter 8, about the relevance of accumulated knowledge concerning reactions of the private economy, was not misplaced.

Ending here our examination of the VAR approach to macroeconomics we borrow from D. Hendry[254] the opposition between "theory dependence" and "sample dependence". Hendry takes the Real-Business-Cycle approach as an example of those which are "theory-driven": the credibility of the models depends much on that of the theory from which they arose and little on data evidence. In contrast, the "data-driven" VAR approach looks for models which closely describe the data, but "may suffer from sample dependence in that accidental and transient data features are embodied as tightly in the model as permanent aspects". In between stands the approach that attempts to merge inference from data with guidelines from economic theory, emphasizing empirical models as reductions which can be congruent with the theory. "All three approaches base their conclusions on a mixture of theory and evidence, but accord very different weight to the components, and often have very different constructions of admissible theory. These major differences in approach reflect genuine difficulties in... economics and are not merely fads."

4.5. The place of calibration in macroeconometrics

In modern macroeconomics calibration happens to be used almost exclusively as a password identifying reference to the empirical work of the Real Business Cycle movement. This is very unfortunate because calibration problems occur in all approaches to macroeconomics. Basically these problems have nothing specific to do with the question of knowing whether the real business cycle theory is true, or whether the research strategy advocated by the RBC movement will eventually prove to be efficient.

[254] D. Hendry, Econometrics and business cycle empirics, *Economic Journal* (November 1995).

The word calibration is actually not so good because it covers three types of operation which are not closely connected with one another. We may speak of calibrating the empirical and simulation work on a given abstract model which is found appropriate for a class of applications; the purpose is to define the correspondence between the concepts of the abstract model and the measured variables used in applications. We may speak of calibrating the value of a parameter appearing in a macroeconomic model when we use external estimates of this parameter, found somewhere in the literature; the purpose is to complete the quantification of the model so that it may be used for simulating the effect on endogenous variables of changes in exogenous variables. We may speak of calibrating a model when we are matching empirical moments, computed from a statistical data base, to those implied by stochastic simulations computed with a model meant to be relevant for predictions involving variables appearing in this base; the purpose is verification or search, verification that the model is acceptable because it provides a good match, or if such is not the case, search for an improved model which will provide a better match.

All three operations are important in empirical macroeconomics. Each requires respect for an appropriate methodology. Since this book does not enter deep into econometrics, it will be short about the issues involved. This section aims at no more than an introduction, hopefully listing and defining the main points to have in mind. It will consider in turn each one of the three operations. But it could not end without briefly voicing indignation against loose practices which, often covered under the name of calibration, disgrace scientists who use them.

4.5.1. *Calibration of the work on a given model*

When a macroeconomist has decided to use an abstract model in order to test a theory on a data set, or to predict some effects of a possible exogenous change in a concrete economy, he or she must exactly define which measurements to use in order to quantify the abstract concepts of the theory. These measurements may be obtained from the national accounts of the country or from some other statistical data. But so much is available in these sources that a selection must be made. More importantly, statisticians, aiming at providing information to a large variety of potential users, give measurements for concepts which intend to be close to actual observation, whereas abstract macroeconomic models simplify in many respects. Any application of the model therefore requires definition of a correspondence matching the concepts of the model to concepts of the data source. Choosing the correspondence is calibrating the empirical work that will use the model.

The importance of the operation in question is not often stressed. Indeed, a frequent practice for authors reporting their results is to relegate in an appendix some brief and incomplete indications on how they obtained the measurements of the variables they are using: to give a full description in this respect might be tedious and might even, in some cases, be quite lengthy[255]. We must therefore be thankful to T. Cooley and E. Prescott who brought to the fore this aspect of calibration in the introductory chapter of the book intended to present the work of the RBC movement[256].

In their definition of the aspect in question they argue as follows: "To go from [our] general framework to quantitative statements about the issues of interest... [requires in particular] to construct a set of measurements that are consistent with [the model]. With enough theory and observations..., we can establish the correspondence between [the model] and the observed data for the US (or some other) economy. As we will demonstrate, establishing this correspondence may well require that we reorganize the data for the US economy in ways that make them consistent with our [model]". The authors then illustrate the calibration procedure on the example of the measurement of the variable "capital", which plays a large part in their simple growth model. They consider the match between this variable and "capital as it is measured and as it is conceptualized in the US National Income and Product Accounts (NIPA)".

The abstract model is meant to cover the whole economy. Capital K in their model must include not only the fixed capital stock of the accounts but also the stock of inventories, the stock of consumer durables, and government capital stock. The two last additions require that output y is not measured exactly by GNP but also includes estimates of the flows of services that households currently draw from consumer durables and government from its capital. Measures of consumption and investment, as they are found in the national accounts, must also be revised in conformity with the above changes.

In their measurement of flows, Cooley and Prescott accept to use the gross concepts rather than the net ones, which would differ by subtraction of capital depreciation. In that they keep the convention which is usually made in uses of national accounts. In order to estimate the flows of services to be added to GNP, they multiply the corresponding capital stock by the sum of the appropriate depreciation rate and net rate of return, the estimation of which they explain.

This calibration of the measurement of variables has implications on ratios to which the model may assign particular significance. For instance, the

[255] This comment is not meant to be a justification of the practice. On the contrary, reproducibility of the empirical works would require full descriptions.
[256] T. Cooley (ed.), *Frontiers of Business Cycle Research* (Princeton University Press, 1995).

model of Cooley and Prescott implies that the exponent of capital in the Cobb–Douglas production function is equal to the share of gross capital income in output. In the time series this share is not constant. But the authors report how they take its average value over the sample period for calibrating the exponent in question.

If another model of the same US economy would have been used, the calibrating conventions would have been more or less different. A more detailed model would have required more conventions, the report of which would have been longer and more tedious, or alternatively less complete.

Existence of an explicit correspondence between the variables of a model and the measurements found in the data sources is not something new, which would have been introduced by the RBC movement. The word calibration was introduced by it in order to designate in particular this matching of the model with the data, but not the concept. Full presentations of structural macroeconometric models all contained a description of the correspondence. For instance, in the 1984 book of R. Fair (op. cit.), 22 pages of Appendix A describe the "constructions of the variables" of the US model that the book analyses. Table A.1 defines the sectors of the model by aggregation of sectors used in the US Flow of Funds Accounts. Table A.2 gives where the raw data used for measurement are to be found in the National Income and Product Accounts as published in the *Survey of Current Business*, in the Flow of Funds Accounts (checks on the consistency with the NIPA being given by Table A.3), in the *Federal Reserve Bulletin*, or in publications of the Bureau of Labor Statistics. Table A.4 presents how all the variables in the model are measured. With few exceptions, this is made in terms of the raw-data variables in Table A.2 or in terms of identities. Six pages describe what was done for the few variables that did not exactly fit in this pattern. These pages also explain the formulas by which the personal income taxes to the federal government, or to the state and local governments, were derived from the amount of the taxable income: the values of the coefficients used in the formulas were obtained from side regressions, which played a similar role to that of the average value calibrating the exponent of capital in Cooley and Prescott.

4.5.2. *Use of external estimates of parameters*

We now turn our attention to the calibration of values of some parameters appearing in a macroeconomic model. To understand the good reasons for such a calibration we have to refer back to the econometric precepts taught by proponents of structural models and to point to the main weakness of those precepts in macroeconomics. But drawing on our microeconometric knowledge in order to calibrate a macroeconomic parameter is not a simple operation, except in an ideal case not often met in practice. A whole research programme is needed if we macroeconomists want to do our empirical work properly in this respect.

(i) We saw in Part 3 of this chapter the context within which an empirical methodology was elaborated at the Cowles Commission more than fifty years ago for application to macroeconomics: a model was specified, which was meant to embody prior knowledge about the phenomena concerned; the relevant additional evidence was given by a data base made of time series of the variables present in the model observed over a sample period in a given geographical area; the specified structure of the model was estimated from the data base by application of an appropriate statistical inference procedure.

As we saw, reservations about this methodology were later expressed arguing that models used in practice were not good representations of prior macroeconomic knowledge. Either, according to C. Sims, models were too specific, so much so as to be “incredible”; or, according to the classical macroeconomics schools, models overlooked what good foundations had to be. Even without sharing either of these two extreme views, a number of economists pointed to substantial revisions to be brought to the specification of existing structural models: their long-run properties had to be better examined and revised, their stochastic dynamic properties had to better agree with those of observed time series, and so on.

But by far the main difficulty with the Cowles Commission methodology as described came from the fact that its results were not accurate enough to meet the claims of proponents of the methodology, and the difficulty remains five decades later, notwithstanding the tremendous increase in the provision of economic statistics and economic data bases. Calibration of some parameters in the models is then viewed as a means to alleviate the difficulty. Before looking at calibration itself, let us sharpen our diagnosis a little bit.

The low accuracy of estimated structural macroeconometric models is particularly detrimental when the answers required from the models concern medium-run effects and/or prediction of what would happen in conditions which seldom occurred in the sample period. The short-run effects of familiar shocks are not so precisely forecast, as we saw in Section 3.5; the models now in existence could certainly benefit from improvements; however, they already give about such effects useful assessments, which are found to be more precise on average than those obtained from alternative techniques. The long-run effects are usually better analysed by more direct arguments than those embodied in structural models. But assessing medium-run effects, after five years, say, is a very serious challenge to macroeconometricians, all the more so as policy makers realized since the 1970s that the medium term needed more attention than had been given to it during the two preceding decades. Finally, finding what would happen after unfamiliar shocks is, of course, intrinsically difficult; we may fear that fits on past time series are particularly imprecise for use in the then needed off-sample extrapolations.

Some improvements might be brought to existing structural models by use of relevant information which either is available in some time series not contained in the commonly used data bases, or could be gathered in new appropriate time series. We naturally think of series of the values taken by indicators of expectations, perceptions and sentiments, revealed by various economic agents about various aspects of their environment. We also think of disequilibrium indicators concerning in particular the degree of capacity utilization, the experience of more or less acute difficulties in obtaining bank loans, and so on. But, in the first place, some such indicators are already present in some macroeconometric models. Second, we should not overestimate the resulting gains in accuracy: for instance, indicators of expectations or even intentions were not found to be so good explanatory variables or predictors in econometric investigations.

Given all that, we may think that today the best sources of available extra information for macroeconomists are to be found in microeconomic data bases. Such bases hardly existed fifty years ago. Many have been produced in the last few decades, often over many thousands of elementary units and containing data on substantial numbers of variables. When econometrically processed this material delivers estimates of parameters, which seem to be significant for macroeconomists. Nobody, except perhaps the most extreme adepts of holistic visions, could deny that macroeconomic research should consider such estimates coming from microeconomic data and take advantage of them in the conception of macroeconomic models. Let us look at the potentials offered by use of this source of information, and also at the problems of its use.

(ii) In order to focus attention on the main point we need not consider a fully-fledged structural model. We may reduce the model to being just a multiple regression, even one with two regressors only, such as:

$$y_t = a_1 x_{1t} + a_2 x_{2t} + b + \varepsilon_t. \tag{323}$$

Besides time series of y_t, x_{1t}, x_{2t}, the econometrician has an estimate $\hat{a}_2$ of a_2, which was obtained from an external source such as a microeconomic data base.

The ideal case for calibration occurs when $\hat{a}_2$ is for sure exactly equal to the true value of a_2 appearing in (323). Then a fit of the time series on

$$y_t - \hat{a}_2 x_{2t} = a_1 x_{1t} + b + \varepsilon_t \tag{324}$$

will give more accurate estimates of a_1 and b than a fit on (323). The theory of multiple regressions shows that, under a wide range of assumptions such as the unobserved disturbances ε_t being independently and identically distributed and being independent of (x_{1t}, x_{2t}), the sampling standard-error of the estimate of a_1 given by the fit on (324) will be all the more lower than that given by the fit on (323) as the correlation between x_{1t} and x_{2t} happens to be higher. Since we know that cases of high colinearity between regressors are frequent in macroeconomics the gain in accuracy, coming from the use of the external source together with the time series, may be quite important in the estimates of parameters of the model. It may be even more important in the policy predictions given by the model if those concern larger shocks on x_{2t} than commonly occurred in the sample period. But the case of the external source delivering a perfect knowledge of the value of a_2 is exceptional. We must now examine the various reasons which make calibration of a_2 by $\hat{a}_2$ less favourable.

First of all, external estimates are subject to *estimation sampling errors*, which may be quite substantial when microeconomic data have been used. Even though the sample size is normally many times larger in microeconomic data bases than in time-series macroeconomic bases, the sampling errors may be relatively high because microeconomic units are much more heterogeneous with respect to economic features than are quarters or years. This means that a large number of regressors characterizing this heterogeneity must be introduced in the microeconometric work besides the regressors in which the macroeconomist is interested.

If the microeconomic estimation error α is the unique reason for $\hat{a}_2$ to differ from the true value of the macroeconomic parameter a_2, calibration of a_2 by $\hat{a}_2$ leads to the following model as underlying the time-series estimation:

$$y_t - \hat{a}_2 x_{2t} = a_1 x_{1t} + b + \varepsilon_t + \alpha x_{2t}. \tag{325}$$

It is usually a very good approximation to assume stochastic independence of ε_t and α. The error term in (325), namely $\varepsilon_t + \alpha x_{2t}$, then has a larger variance than ε_t does (it is also heteroscedastic, which makes the efficient estimation procedure more complex than a simple regression). Hence, the gain in accuracy coming from the use of the external source is smaller than in the ideal case.

But there are also reasons for fearing that, as an estimate of the macroeconomic parameter a_2, the value of $\hat{a}_2$ is subject to *systematic errors*. A frequent reason concerns the case of dynamic macroeconomic models when the external source is a representative cross-section of individual units, all observed in a single period. In order to understand the problem, let us neglect the heterogeneity of individual units ($i = 1, 2, \ldots, n$) and assume that their behaviour is well described by the following model:

$$y_{it} = a_1 x_{1t} + a_{20} x_{2it} + a_{21} x_{2i,t-1} + b + \varepsilon_{it}. \tag{326}$$

We note that, in this particular case used here for convenience, the same exogenous variable x_{1t} is assumed to enter the determination of all individual y_{it} (like the same price level for all households).

We also note that a lag occurs in part of the effect of the individual exogenous variable x_{2it}. The microeconomic model (326) nicely aggregates into the following macroeconomic model:

$$\bar{y}_t = a_1 x_{1t} + a_{20} \bar{x}_{2t} + a_{21} \bar{x}_{2,t-1} + b + \bar{\varepsilon}_t, \tag{327}$$

where the macroeconomic variables $\bar{y}_t$, $\bar{x}_{2t}$ and the disturbance $\bar{\varepsilon}_t$ are simple averages of the corresponding microeconomic variables and disturbance.

We assume that the external cross-section contains only the individual values $y_{i\theta}$ and $x_{2i\theta}$ for a particular period θ (the value of $x_{1\theta}$ being otherwise known, but this does not much matter). The external estimate $\hat{a}_2$ has been obtained by a fit of the microeconomic data on:

$$y_{it} = a_2 x_{2it} + c + \varepsilon_{it}. \tag{328}$$

We easily see that $\hat{a}_2$ is in general a biased estimate of both the short-run parameter a_{20} and the long-run parameter $a_{20} + a_{21}$. If a regression was used for the calculus of $\hat{a}_2$ and if the standard hypotheses of regression theory apply, one easily proves that the mathematical expectation of $\hat{a}_2$ is equal to:

$$E(\hat{a}_2) = a_{20} + \frac{\sigma_1}{\sigma_0} r_{01} a_{21}, \tag{329}$$

where σ_1 and σ_0 are the respective standard deviations of the microeconomic variables $x_{2i,\theta-1}$ and $x_{2i\theta}$ over the population of individual units, r_{01} being the correlation coefficient between the same two variables. In comparison with a_{20} and $a_{20} + a_{21}$, the estimation $\hat{a}_2$ is hybrid. Most often in cases of that sort, the correlation r_{01} is positive, and even fairly high. So the bias of $\hat{a}_2$ with respect to the long-run coefficient (here $a_{20} + a_{21}$) may be fairly small; calibration of $a_{20} + a_{21}$ by $\hat{a}_2$ leads, however, to underestimate the long-run effect.

Such biases are avoided if the data base contains not only contemporaneous but also lagged data on the variables of interest, if for instance here the values of $x_{2i,\theta-1}$ are recorded so that two values $\hat{a}_{20}$ and $\hat{a}_{21}$ may be estimated and serve for calibration of respectively a_{20} and a_{21} in the work on macroeconomic time series. The difficulty is still better avoided if the data base comes from a representative panel of households or firms followed on a sufficient number of periods, so that the risk of the sample period being special is minimized. Indeed, the fact that macroeconomic models are usually dynamic, sometimes assuming even a fairly complex dynamics, is one of the reasons why panels of economic data are gathered.

The risk of systematic errors in the calibration of macroeconomic parameters by a direct transfer of estimates obtained from microeconomic data appears still more troublesome when we think about the intricacies of aggregation, to which we shall turn in a moment. Before doing so we still need to insist, and to stress the pervasiveness of a consideration that may not appear on such a simple model as that of Equation (323).

(iii) Most parameters for which calibration is entertained measure intensity of some differential effect. Each one concerns in its model a function of usually several variables. In other words, each parameter has the nature of a partial derivative or elasticity of a function. It may be like a first-order derivative or a higher-order derivative (such as an elasticity of substitution). A very frequent source of bias in calibration arises because all macroeconomic models ignore, in the specification of the functions they contain, some of the many variables that might play a part in the determination of the dependent variable. This omission, which affects the meaning of the parameter, may not be serious for intended uses of the macroeconomic model, but it is often not consistent with the definition of the corresponding parameter in the microeconometric work which gives the calibrating value.

In other words, *quite often the microeconometric source estimates a different parameter, i.e. a different partial derivative, from the one it is calibrating in the macroeconomic model*. Comparison between Equations (327) and (328) illustrates the problem. But, whereas with the simple case studied above it was possible to have an idea of the bias with respect to either a_{20} or $a_{20} + a_{21}$ by reference to Equation (329), the bias is in general much more difficult to appraise: *de facto* it is commonly ignored. This point is righty made by L. Hansen and J. Heckman[257] who write in particular: “Microeconomic studies rarely estimate models that can be directly applied to the aggregates used in the real business cycle theory”. These authors also stress that different microeconometric studies commonly estimate different parameters and that it is difficult to draw, from them all, the values which would be appropriate for the economic environment discussed in a particular RBC model.

Such a serious reservation does not mean that we should systematically forgo calibration of parameters in macroeconomic models from microecono-

[257] L. Hansen and J. Heckman, The empirical foundations of calibration, *Journal of Economic Perspectives* (Winter 1996).

metric results. But the operation is exposed to risks, which should be kept in mind and ought to be studied whenever the value of a calibrated macroparameter plays a crucial role in the applications of the model. This recommendation to be cautious is reinforced when we think about aggregation.

(iv) *Aggregation biases are particularly difficult to evaluate* because they arise from multifarious causes, which are moreover often specific to the phenomena playing the dominant part in particular macroeconomic applications. Detecting the presence of such errors requires a wide economic and statistical culture, assisted by a shrewd intuition.

We may indeed reflect back on the various cases of aggregation discussed in this book, which cannot however claim to offer a picture spanning the whole spectrum of relevant cases. In the typical situation we needed introduce a more elaborate and specific modelling than had been defined in the general formalization of Sections 2.1 to 2.5 of Chapter 2. We realized it immediately after, when discussing the life-cycle effects on aggregate saving or the dependence of aggregate saving on the interest rate. We saw it again in Chapter 4 where in Part 6 we, however, limited attention to the aggregation of production functions. This was still the case in the otherwise simple analysis of aggregation over markets experiencing different degrees of tension (Sections 4.8 of Chapter 7, 8.7 of Chapter 8 and 4.1 of Chapter 9).

At these various occasions we did not extensively explore the effects of microeconomic heterogeneity. But we saw in Section 6.2 of Chapter 4 a case suggesting that, because of this heterogeneity, aggregation biases may be particularly serious for parameters which have the nature of second-order rather than first-order derivatives. Special caution seems to be required for calibrating from microeconomic sources parameters such as macroeconomic elasticities of substitution.

(v) In conclusion, we cannot object in principle to calibration of some parameters in our macroeconomic models: the potential gain in accuracy is too substantial, relatively to what can be otherwise achieved. We may even anticipate that the practice of parameter calibration will spread. Macroeconomists have indeed much to learn from the information content of microeconomic data bases, especially when they are panels of microeconomic data, which are available in increasing number. But we must keep in mind that calibration of parameters is a risky operation, which may introduce systematic errors in macroeconomic models.

Writers of surveys of econometric results obtained for the characterization of some aspect of behaviours are therefore providing an important service to macroeconomists. We had occasions to quote a number of such surveys, in particular two extensive ones due respectively to Angus Deaton, about consumption and saving in Chapter 2, and to Daniel Hamermesh, about the demand for labour in Chapter 4.

In order to minimize the risk of systematic errors, users of such surveys for calibration purpose have, first, to reflect in each case on what is the exact definition of the parameter they want to calibrate, second, to search within the survey whichever results concern the same definition. Knowing that their surveys will be so used, authors should carefully specify the definitions corresponding to the various results reported, which usually come from econometric works referring to different definitions, in particular because the sets of omitted variables differed.

4.5.3. *Matching observed moments: verification and search*

The third operation covered over the name of calibration in the literature about real business cycles and its successor consists in matching results given by a model to observed moments such as the variance of the main variables and their contemporaneous or lagged correlations with GDP. The purpose of the operation is well suggested by the two following sentences drawn from T. Cooley and E. Prescott (1995, op. cit.): "Judged on the dimension of the composition of output and its comovements, the match between the model economy and the observed data for the US economy is pretty good but not perfect... The failures of the model economy tell us there are important margins along which decisions are made that have not been captured in this simple world...". The first sentence expresses the satisfaction to see that the fit is "pretty good", which would provide a verification that the simple "model economy" already achieves a good job. The second sentence announces the subsequent chapters of the book which will intend to capture other types of decision and so search for a better fit.

(i) We were rather critical, at the end of Section 2.3, about the type of *verification* of the real-business-cycle theory that was so given. We stated that the test had low power and that other theories could, just as well, pass the same test. Without denying at all this earlier negative assessment, we must now take *a somewhat more sympathetic view of the calibration approach* under discussion, in comparison with the alternatives really available to econometricians. For so doing we must briefly refer to the fundamentals of statistical inference.

Quite naturally econometricians adopted fifty years ago the methodology of classical mathematical statistics: inferences from new data were assumed to take place within the framework of a stochastic model gathering what was already known (this model is often called "the maintained hypothesis"). For the study of the relative performances of alternative estimation procedures (various "estimators"), this methodology is most often found appropriate. But it was often questioned when tests of alternative theories were at stake. The main concern was a mismatch between the way in which the problem was posed in the teaching of mathematical statistics and the way it was perceived to be actually posed in research about economic theories.

Mathematical statistics looks at the test of an hypothesis H_0 which is a particularization of maintained hypothesis H, i.e. of the model assumed to be true. Hypothesis H_0 will be rejected if, given H and the data, it is unlikely to be true. Research will then be geared away from H_0, towards the test of any other particularization of H that scientists may conjecture to be true. We thus have a vision of scientific advance resulting from a quest for "falsifications", a vision worked out by the philosopher Karl Popper[258].

Insatisfaction about this way of posing the progress of knowledge in economics was felt by a number of those who worked on economic methodology. It was best expressed for our purpose here by E. Leamer[259] who wrote in the introduction to his book: "In searching for a model of nonexperimental inference, we may easily discard the textbook version of classical inference. It makes implicit use of the following unacceptable specification axiom" (the maintained hypothesis as usually specified). Leamer explains that specification search is actually data mining led by judgement and purpose, the latter varying over a wide range from improving an existing model to constructing a "fruitful model". For such operations which are intrinsically difficult to formalize in general terms, Leamer claims that the Bayesian view of statistical inference yields useful insights.

We must recognize that, in this book, we did not explicitly refer before this point to the classical framework of statistical tests, but it was implicit each time we spoke about the significance of an effect estimated in an econometric fit. At other times referring to it would not even have been helpful. For instance, considering in Subsection 3.5.4 the evaluation of structural macroeconometric models, we mainly discussed the predictive accuracy of these models in comparison with what could be obtained by VAR extrapolations. This seems to imply that verification of this accuracy was essential, that improvement in it would have been a legitimate purpose for research and that, for so doing, no maintained hypothesis was required[260].

[258] For a presentation and its application to economics, see M. Blaug, *The Methodology of Economics* (Cambridge University Press, 1980).

[259] E. Leamer, *Specification Search* (John Wiley, New York, 1978).

[260] Remember, however, the end of Section 3.5 where we examined an econometric test devised and applied by R. Fair and R. Shiller in order to know whether a structural model was carrying

The practice of calibration in its third sense by economists of the RBC movement is, of course, not exactly what we then did. It rather stops at inspecting whether there is a good fit between the values predicted by the model for a few statistical indicators and the values observed for the same indicators in the economy. Attempts at precisely evaluating the quality of the fit are quite exceptional. As far as this author knows, no systematic comparison was ever made with the fit that could have been obtained, on the same indicators, by VARs or structural models. Again the verification commonly offered by RBC authors is not stringent enough.

(ii) Let us discuss this point and look at some applications of econometric theory to RBC models in order, in particular, to precisely evaluate the *goodness of fit* to the data[261].

Quite naturally members of the RBC movement compare to the data the results of simulations made with their models. This approach is commendable and not special to the movement: it is commonly used for the evaluation of structural macroeconometric models since the initial work of L. Klein and A. Goldberger (1955, op. cit.). We saw this role of the approach particularly in Section 3.5. Introducing the neologism "computational experiment", as is done by some proponents of calibrations[262], serves no useful purpose. It can only be confusing when it conveys the notion of a fundamentally new methodology.

We noted that the statistical indicators selected by calibrators for evaluating the goodness of fit were essentially second order moments: variances and correlation coefficients. We may see that this is not special either to the RBC procedures. When evaluating the predictive accuracy of structural models we also selected first and second-order moments of prediction errors (the two kinds of moment being associated in the computation of root-mean-squares of these errors). The implicit assumption was then that the loss resulting from the errors would be well represented in most applications by a quadratic function: the expected loss would then be a function of first and second-order moments only (this would hold exactly in a linear economy, approximately if such an economy provided a sufficient reference). We shall not quibble over this feature, but rather consider two questions: how to rigorously evaluate the goodness of fit? Which moments of the data should we particularly match?

useful information ignored by a VAR and so had a significantly higher predictive potential.

261 For the presentation of other applications, concerning also calibration of parameters, see A. Gregory and G. Smith, Statistical aspects of calibration in macroeconomics, in: G. Maddala, C. Rao and H. Vinod (eds.), *Handbook of Statistics*, Vol. 11 (Elsevier, 1993), pp. 703–719.

262 F. Kydland and E. Prescott, The computational experiment: an econometric tool, *Journal of Economic Perspectives* (Winter 1996).

(iii) An interesting study in this respect is due to *an article of L. Christiano and M. Eichenbaum* to which we referred in Section 2.4 about the introduction of shocks to both aggregate demand and the technology[263]. Presenting their econometric methodology the authors write the following sentences, which are worth a full quotation here:

"Much of the existing RBC literature makes little use of formal econometric methods, either when model parameter values are selected or when the fully parametrized model is compared with the data. Instead, the RBC literature tends to use a variety of informal techniques, often referred to as *calibration*. In contrast, we use a version of Lars Peter Hansen's generalized method-of-moments (GMM) procedure at both stages of the analysis[264]. Our estimation criterion is set up so that, in effect, estimated parameter values equate model and sample first moments of the data. It turns out that these values are very similar to the values used in existing RBC studies. An important advantage of our GMM procedures, however, is that they let us quantify the degree of uncertainty in our estimates of the model's parameters. This turns out to be an important ingredient of our model-evaluation techniques".

Let us look at a few details of the procedure as applied in the article and at a few results. *Concerning estimation of the parameters* the main point is to note that the authors do not appeal to any external source, which would have given "calibrated values". They infer the estimates from the same aggregate time series data base that is also used later for evaluating the model (US quarterly series from the middle of 1955 to the end of 1984). As is reported in the above quotation, the estimators are computed from the arithmetic sample means of some appropriate quantities. For instance, the estimated rate of capital depreciation per quarter is the sample average of the quarterly depreciation rates as obtained from series of capital and gross investment. The estimated exponent of capital in the Cobb–Douglas production function is derived from the Euler equation implying equality between, on the one hand, π_t equal to one *plus* the net marginal productivity of capital in quarter t and, on the other hand, ρ_t equal to the marginal rate of substitution between the utility of consumption in quarter t and the discounted utility of consumption in quarter $t+1$ (this equation, implied by the model, appeared for instance as Equation (74) in Chapter 6 of this book): the capital exponent is estimated as being such as to equate to 1 the sample mean of the quarterly ratios π_t/ρ_t (we note in passing that the expression for ρ_t assumes a logarithmic utility function and involves the factor

[263] L. Christiano and M. Eichenbaum, Current real-business-cycle theories and aggregate labor market fluctuations, *American Economic Review* (June 1992).

[264] L. Hansen, Large sample properties of generalized method of moments estimators, *Econometrica* (July 1982).

β discounting utilities; this factor was not estimated but set "so as to imply a 3-percent annual subjective discount rate", but choosing 5-percent, say, would not have much changed the results). Other parameters were estimated with reference to similarly appropriate sample means in each case.

Estimation concerns in particular the standard errors of two unobserved independent white noise processes generating the shocks to respectively the technology and government demand, these two processes being responsible for sampling fluctuations. The asymptotic theory of GMM estimators provides formulas for assessing the standard errors of all the parameter estimates, assuming the model to be true. In the particular case these latter asymptotic standard errors turn out to be almost all quite small. This work proves that estimation of RBC models by rigorously established econometric methods is feasible.

(iv) Christiano and Eichenbaum want to also *assess the validity of the model*. Indeed, in their work as well as in earlier RBC works, the fit of statistical moments involving the labour market is poor. In particular the model predicts a lower relative volatility of the labour input (in man-hours) than is observed in the data (volatility relative to that of output or even labour productivity, and measured in terms of variances of time-series of these variables). The model predicts a strong positive correlation between labour productivity and labour input, whereas just a weak correlation is found in the data. Such counterfactual evidence was often stressed in discussions about the validity of RBC models, as we saw when dealing with these models in Part 2 of the Chapter. But, can we assess the significance of the evidence?

Facing the question with their data and using the classical statistical approach, Christiano and Eichenbaum have to reflect about the maintained hypothesis (H) within which the validity of their model (H_0) will be judged. The asymptotic theory of GMM estimators suggests an hypothesis H and a test of H_0 which should be appropriate for the purpose. Let us understand how.

The first step in this theory is to show that estimators are "consistent": their values would converge to the true values if the sample size would increase to infinity. This holds because of the choice of the system of equations on sample means which defines the estimators: this system, let us call it S_1, applies when sample means are replaced by mathematical expectations of the corresponding variables, moreover sample means converge to precisely these mathematical expectations. The model H_0 implies such a system of equations on mathematical expectations, actually 7 equations corresponding to those used for determination of the estimated values of the 7 parameters of the Christiano–Eichenbaum model.

This model of course implies much more than S_1. For instance, it implies how the true values of second-order theoretical moments[265] of the endogenous variables can be computed from the true values of the parameters. The function so giving the moments from the parameters are used for the computation of the values of the moment "predicted by the model" from the estimated values of the parameters. Let us call F the system of these functions and ω_0^* the vector of second-order moments so estimated through F from the estimated values of the parameters.

An alternative method for estimating the second-order theoretical moments is, of course, to use corresponding sample moments. Let us call ω^* the vector of estimates so obtained. Although far-fetched in appearance, it will be convenient here to see the determination of ω^* as a GMM estimator using the system S_2 of identities defining the second-order theoretical moments from the mathematical expectations of quadratic functions of the variables. We then have two estimators of second order moments, leading respectively to ω_0^* and ω^*. If H_0 is true both estimators are consistent, converging to the same theoretical moments. So we may suspect that, if ω_0^* and ω^* are far apart, H_0 is unlikely to be true. In order to define a test of H_0 we just need to define a distance from ω_0^* to ω^* and to find out when this distance will be too large for H_0 to be maintained. Clearly to make sense of the approach we need the vector ω^* of sample moments to be a consistent estimate of the vector of the theoretical moments considered. Since the maintained hypothesis H has not to insure more than this consistency, we may say that H is indeed large.

We do not want to go more precisely here into the asymptotic theory of GMM estimators and into the characterization of the asymptotic probability distribution that $\omega_0^* - \omega^*$ would have if H_0 was true. We understand how this distribution can be obtained by comparison of two GMM estimators: ω_0^* using S_1 for the estimation of parameters and then F, ω^* being part of a larger vector in which both the parameters and the theoretical moments would be estimated using the system made of S_1 and S_2, but neglecting the relation F between parameters and theoretical moments. The statistical test then consists in deciding to reject H_0 if the value found for the distance between ω_0^* and ω^* has a too low asymptotic probability to be reached when H_0 applies.

The test performed by Christiano and Eichenbaum is not exactly the one we just sketched because they did not consider a distance between ω_0^* and ω^* but rather a distance between $g(\omega_0^*)$ and $g(\omega^*)$, the function g mapping second-order moments into the vector of two indicators measuring respectively the

[265] Here second-order "theoretical moments", in contrast with second-order "sample moments", refer to the mathematical expectations of centered quadratic functions of the variables, rather than to the sample means of the same functions.

volatility of labour input, relative to that of labour productivity, and the correlation between labour input and labour productivity (these two indicators were indeed those for which there was an apparent mismatch). This difference does not essentially change the methodology since the test refers to an asymptotic probability distribution, which has simply to be worked out from the asymptotic probability distribution of the joint vector (ω_0^*, ω^*).

The results found by this application of classical econometric theory confirmed the initial suspicion. Christiano and Eichenbaum dealt with four versions of their model and two data sets, differing by the series chosen for the measurement of labour input, hence also labour productivity. The asymptotic probability of reaching such a high value of the distance between $g(\omega_0^*)$ and $g(\omega^*)$ as was observed was practically zero in six cases, 0.07% in one case and 18% in the last case.

(v) *Were the most revealing moments chosen* for evaluating the goodness of fit? We may wonder. Indeed, we remember the study reported by T. Cogley and J. Nason (1995), an article which we examined in Subsection 2.5.2. That study did not aim at devising a particular goodness-of-fit test. Its purpose was rather to demonstrate that the RBC models failed to match the dynamic features of actual macroeconomic time-series. The demonstration was quite persuasive. It showed in particular that the spectrum implied by the model of Christiano and Eichenbaum for the growth rate of output was fairly flat whereas the spectrum actually observed had high values for frequencies in the range from two to seven year per cycle. Thus Christiano and Eichenbaum, who tested the match when looking at contemporaneous second-order moments, did not pay attention to the mismatch which appeared when Cogley and Nason looked at what could be inferred from the series of lagged moments of the growth rate.

Notice that this problem concerns the dangers to which verification by calibration is exposed, but just as well those to which falsification by statistical tests is also exposed. The possibility always exists that research does not look in the direction which would be most revealing: for invalidating verification in the first case, and for exhibiting falsification in the second. In a sense the classical theory of statistical tests draws attention to the problem, because it teaches that different tests commonly have different powers with respect to given deviations from H_0 and that the ranking of two tests according to their powers often changes when focus changes from one type of deviation to another. Statistical theory even brings a number of general results about tests which are efficient for testing H_0 against a given alternative H_1 or, in favourable cases, about tests which are uniformly efficient against all hypotheses other than H_0. But in the latter case, the efficiency property then asserted is conditional on what is the maintained hypothesis H, and again some types of deviation from H_0 may be so ruled out.

It is conceivable that, and indeed in some special cases it happens to be true that, given H, the uniformly efficient test of H_0 exists and implies consideration of a particular statistical moment of the data. The question "which moment to match?" then has a rigorous answer. Or still, it may happen that a particular set of moments provides a sufficient statistics for testing H_0 within H (by definition of sufficiency it means that the set conveys all that information contained in the data which is relevant for this testing). We then so know which moments we ought to match. Thus in some cases, questions naturally raised by calibration in its third sense have neat answers. Unfortunately, as soon as we have a sufficiently broad concept of H, these cases appear exceptional for various reasons (including that the world is not Gaussian, even approximately, whereas the Gaussian assumption if made in most known results about uniformly efficient tests or sufficient statistics).

(vi) Let us conclude this long subsection with *a few propositions and comments*.

The concern for verification exhibited by calibrators is commendable. After all, when taken seriously, it does not much differ from the classical concern for testing.

Econometric theory brings some lessons about the way in which the concern for verification can best be met. At the bottom, however, good response requires not only rigour and familiarity with econometric theory but also judgement so as to detect what could seriously invalidate a given model.

Matching first and second-order moments provides an interesting approach to verification. But there are potentially many such matches to make, rather than the few to which attention was most commonly geared. We may even anticipate that this large family of potential matches will not always be found to suffice[266].

Christiano and Eichenbaum used GMM for the estimation of parameters of their model and the test of its match with the data. Other fairly well charted methods are also available in econometric theory for the same purpose and are indeed used. Moreover, as a careful reader of the foregoing pages has certainly realized, moments to which reference is made in the definition of the

[266] K. Wallis (1995, op. cit.) points to the fact that stochastic simulations from a macroeconomic model can be used to estimate the probability of occurrence of certain economic events. Such probabilities could be matched against actual frequencies of the same events, conditional on the values of exogenous variables. Similarly macroeconomists who identified non-linearities in the behaviour of important time-series should like to consider higher-order moments or specific statistics of different kinds.

Generalized Method of Moments may differ from the first and second-order moments of the data series (the moment for the estimation of the capital exponent in Christiano–Eichenbaum is the mean of a function of the data which is neither linear nor quadratic). Overall, the notion of an intimate interlinkage between the three operations of validation by calibrators, matching observed moments, and application of GMM, would probably introduce more confusion than insight.

4.5.4. *The ethics of calibration*

Generally speaking, ethics is important when the proper rules of behaviour are not obvious or when they are not easily enforced. If ethics matter in empirical economics it is because of the second reason. The scientific ethics of objectivity and faithful transmission of knowledge has clear implications in this field as elsewhere, but breaches of ethical rules may remain longer unnoticed because, in many cases, full reports of the research process from which the empirical results emerged would be too tedious and/or replication of the research does not look worth-while.

Given the non-experimental nature of most research in our discipline and the complexity of the world it is studying, full reports would have to give many more details and references to sources than readers care for, ready as they are to take results on trust. Moreover economists know that empirical discoveries are quite exceptional in economics where knowledge of the facts comes rather from accumulation of evidences and progressive advances in learning from them[267]. Replication of any particular research naturally appears less interesting than if it had led to a discovery. Given this situation, the profession ought to develop appropriate ethical standards, together with sanctions hurting the reputation of those who would have been found not complying.

Clearly, the RBC school did not pay attention to this consideration when contributing to the diffusion of loose practices about how to give references to results of empirical economics. L. Hansen and J. Heckman (1996, op. cit.) make this point forcefully. In particular they quote a sentence which was read with indignation by many others, who well knew what to think about the accuracy of available estimates of elasticities of substitution. These readers could only interpret the sentence as motivated by the wish to oversell a particular line of research. Similarly appealing to "common calibrators knowledge", as

[267] I tried to explain why this is so in E. Malinvaud, Why economists do not make discoveries, in: H. Wolf (ed.), *Contemporary Economic Issues: Macroeconomics and Finance*, Proceedings of the eleventh world congress of the International Economic Association, Vol. 5 (Macmillan, London, 1998).

was done on occasion, looks rather like contributing to the misbehaviour of a number of calibrators.

As we saw in Subsection 4.5.2, calibrating some parameters by drawing from results of microeconometric studies ought to improve the reliability of macroeconomic models. But the operation is exposed to risks which should be signalled and discussed. This too should belong to the ethics of calibration.

4.6. Policy making with macroeconomic tools

Our book on macroeconomic knowledge and analysis cannot close without inspecting how the use of macroeconomics for policy is evaluated, and perceived to operate, at the end of the twentieth century. This use was a dominant motivation all through the development of the discipline, and still is now. But major changes occurred in how policy makers and macroeconomists viewed it. At many places we dealt with policy questions and referred, often implicitly, to how they were posed to macroeconomists, then expected to provide relevant answers. We did not dwell on what was perhaps specific to a particular time or context. Neither did we embark into an history of the changing dialogue between politicians and economists. Readers understood such was not our subject. But they are entitled to request from the author clarifications on how the dialogue now functions. Before an explanation of the way in which this will be attempted here, let us briefly look back at the places where we came close to discussing the relations between policy makers and macroeconomists.

Right at the start of our study of the economic system as a whole we found it convenient to consider the normative theory of economic growth, with its typical object "the choice of a growth strategy that would be centrally decided" (first page of Chapter 5). We immediately stressed that the theory in question had other purposes than contributing to the methodology of central planning, which no longer seemed to be directly relevant. Policy issues were at the forefront later in the chapter with Part 5, concerning the volume of public investment, and Part 6, concerning long-term financial strategies as analysed by the "money and growth" theoretical literature.

Again from the start of Chapter 7 policy issues were announced as an important motive to study the short-term equilibrium. The objective of "throwing light on short-term economic policy" was asserted. The chapter claimed to show "how instruments can be used to control the economy to a certain degree" (p. 789). In Chapter 8 also policy issues played a major role, with not only, in Parts 4 and 7, the dynamic study of effects already shown in Chapter 7 to result from policies, but also the problem of making policies credible (Part 3) and the debate about the effectiveness of monetary policy (Part 5). The present chapter, although centred mainly on business-cycle facts, seriously examined structural macroeconometric models and how they were used for policy analysis within an ideal framework described in Section 3.6 but now seriously contested, as we shall see. The last Section 3.8 in the part devoted to structural models provided a first discussion and evaluation of such models for policy purposes.

We have now to resume the discussion, again wondering about the relevance of these various earlier pieces for actual situations which more or less differ from what is assumed in the ideal framework. We shall therefore begin, in a first subsection, with a short description of institutions shaping the actual structures within which policy decisions are prepared and with a brief survey of what participants in these structures reported about their work. The second subsection will aim at providing a closer view of the interactions between models and policies than was provided by the Frisch–Tinbergen theory of economic policy. The third subsection will comment on the lessons to be learned from the model-policy interaction.

4.6.1. *Economic policy making in practice*

In actual institutions the casting which assigns roles to different kinds of actors is pretty clear: it involves deciders, economic analysts and, in between, economic advisers. Deciders are the central government, with its executive and legislative branches, and the central bank, which is now independent in many countries. The autonomous part played by local governments in consciously acting on the macroeconomic situation seems to be limited to employment in their region and the management of their public debt; it is taken as minor and neglected here. In central government the ministry in charge of economy and finance (the Treasury, say) has the main direct responsibility for the fiscal (and budgetary) policy. In central banks monetary policy decisions are usually taken by a special committee, like the Governing Council of the European Central Bank in the Euroland or the Federal Open Market Committee in the US.

It is more difficult to precisely circumscribe the constellation of economic analysts, the studies of whom may significantly influence macroeconomic policy decisions. It certainly contains teams working in the administration of the Treasury or the central bank. But non-governmental institutes such as OFCE in France (Office Français des Conjonctures Economiques) or NIESR in the UK (National Institute of Economic and Social Research) undoubtedly also play a part through their publications. This can also be said of some academic institutions and even of some private firms such as DRI in the US (Data Resources Incorporated).

When considering the above short list we are naturally thinking of offices or institutions employing mainly macroeconometricians. But the definition would then be too narrow. Not late after a major change in the trends of macroeconomic thinking occurred, interest was shown in governmental circles. This happened in the 1950s with the diffusion of Keynesianism, in the 1970s with the emergence of supply side macroeconomics, in the 1980s with the renewed interest for monetarism and classical economics. This means that the outcome of a large part of theoretical research may also be perceived to be relevant. So, the meaning given here to the phrase "economic analyst" has on occasion to be quite broad.

Most interesting from our present viewpoint are economic advisers, who have frequent and direct access to the deciders for whom they are working, and who are supposed to be familiar with the studies produced by economic analysts. These may be just individuals or small teams such the three members of the Council of Economic Advisers in the US. Nowadays most of these advisers have a staff directly helping them in performing their function. Wondering about what this function is, and was, supposed to be will lead us to better understand how in practice the conception of macroeconomic policy making evolved and where it now stands. While doing so, we shall keep in mind that the same advising function is also practised by groups or even institutions which have less frequent access to policy makers but work mainly on policy issues. We may take as example the Council of experts (Sachverständigenrat), made of five persons with academic affiliation sometimes called "the wise men", in Germany. The OECD plays a similar role at the international level.

At this point it is helpful to place our examination of policy advising in an historical perspective. Economic advisers were already active in the nineteenth century. Two French economists became famous precisely because of that: Michel Chevalier, trained as an engineer, an adept of the industrialist Saint-Simonian movement, but progressively becoming a proponent of economic liberalism; Jean Courcelle-Seneuil, who now appears as the first good example of a foreign economic expert because of his role in Chile (both have entries in the *Palgrave Dictionary*). It seems that, in the late nineteenth and early twentieth centuries, economic advisers were seen as experts in the rules of the market economy and as guardians of economic orthodoxy: they were supposed to watch politicians and to protect them against their inclination to interfere in the markets; they had to exhibit the damaging long-run effects that would result from interventions; they had to propose ways for dismantling institutions responsible (in modern parlance) for market rigidities and rent-seeking behaviour.

This was then a quite different function from the policy engineering promoted in the 1950s by J. Tinbergen and others. But a sharp opposition would be misleading as we are going to see. In the first place, economic theory provided in those early days hardly any other framework for the study of active economic policies than one in which they would be detrimental beyond the domain of building public equipments and running public services. Macroeconomics did not really exist as a discipline, as was reminded by A. Robinson, one of those who worked in England with Keynes for policy advising: "There was no macro-economics [in the 1920s]... We lived in a world that fluctuated, which we believed to be uncontrollable but somehow self-righting"[268].

Notwithstanding the radical changes that followed the Great Depression and the advent of Keynesianism, microeconomic advices remained an important task for those working close to government and, at least in the US, they did not much deviate from the line of earlier decades. This appears in a short Symposium published on the occasion of the fiftieth anniversary of the Council of Economic Advisers[269]. On pages 13 and 18, H. Stein writes: "The CEA has been a strong advocate of deregulation, and where deregulation was not possible, it has been a strong advocate of the use of prices as the preferred means of achieving the object of regulation, as in the environmental case". "Almost all economic observers agree that microeconomic advice from the CEA has been important and successful". From a detailed examination of microeconomic positions taken by the CEA by C. Schultze I cannot resist the temptation of selecting at pages 28 and 31 the two following quotes: "I am confident that a large majority of former CEA members would identify the Office of Management and Budget as a consistent CEA ally on microeconomic issues, while the Departments of Labor, Commerce and Agriculture are common adversaries". "Top-level economic advisers ought to see their role in the political arena... Instead of offering advice that seeks to balance economic insights, institutional views, political costs and other considerations, CEA members should see themselves as *partisan advocates of the efficient solution*... If economic advisers do not speak for efficiency, who will?"

On macroeconomic issues the dominant attitudes among economic advisers changed several times during this century. But these attitudes were, all through, more pragmatic and eclectic than a literal reading of concomitant theoretical writings about macroeconomic policy would imply. Although there are exceptions, most advisers like to insist on the idea that they are "moderate" and

[268] A. Robinson, Impressions of Maynard Keynes, in: D. Worswick and J. Trevithick (eds.), *Keynes and the Modern World* (Cambridge University Press, 1983).

[269] Fifty years of the Council of Economic Advisers, *Journal of Economic Perspectives* (Summer 1996).

"practical". Such a declared stand must be viewed in a context in which deciders do not always follow the advices given and tend to add a substantial dose of practicality. Analysing in the CEA Symposium the impact of Keynesianism on the US economic policy, B. De Long substantiates a simple conclusion: "The largest shift in policy marked by the 1946 Employment Act, [which instituted CEA], is the post-World War II practice of allowing the government's automatic stabilizers to function. Not since the Great Depression have mainstream legislators or opinion leaders called for fiscal austerity in the midst of recession. As a result, the federal government's budget exhibits substantial cyclical variation, sliding into deeper deficit in recessions and moving toward balance or into surplus as the economy expands... The commitment to macroeconomic management has checked pressures for fiscal "prudence" that would otherwise have led to *procyclical* discretionary fiscal policy". De Long further observes that the same shift, introduced in 1946 in the understanding of the macroeconomic responsibility of fiscal policy, also concerned monetary policy which since then and up to now was conducted in the US "with an eye not just on price stability but on a whole host of other factors as well".

The hindsights just given about the function and positions of economic advisers could of course be enriched by references to recent experience in other countries. But the picture of the relations between advisers and deciders would not differ much from the one suggested by the above quotations. How about the relations between economic advisers and academic economists? Such is the main question raised in a recent book edited by four Dutch economists[270].

In their introduction the editors point to the fact that communication between academic research and policy advising is subject to frictions. At times and places it functions well, but it may also be lacking "where academics are disinterested and policy-makers ignorant". The editors explain it by the fact that economists play different roles which require different skills. In the concluding chapter they distinguish four markets for the services delivered by economists: fundamental research, applied research, policy advice, and enlightenment of the public on economic issues. "When economics had not yet established itself as a profession, an economist was a multi-market economist. Now that economics has turned into an industry, each economist specializes in a certain mix of techniques and typically focuses on a specific market" (p. 189).

There is, of course, much to be said for this division-of-labour argument. But it should not be taken too literally, neither as a positive account of the facts

[270] P. van Bergeijk, L. Bovenberg, E. van Damme and J. van Sinderen, *Economic Science and Practice: The Roles of Academic Economists and Policy-Makers* (Edward Elgar, Cheltenham, 1997).

nor as a recommendation, particularly if we limit attention to macroeconomics. Many of those who served as policy advisers also worked for research, even fundamental research. There are many ways in which research can advance macroeconomic knowledge; some of them have little or nothing to do with policies; but lacking a concern for policies, and for what they can achieve, would seriously handicap anybody who would like to contribute to general macroeconomic theories of the short run or business fluctuations.

To summarize, let us simply say that an economic adviser is an intermediary between those supposed to know and those supposed to act on the policy front. In order to well perform his or her function the adviser must, like other intermediaries, understand both sides: what scientists can deliver and what deciders really want. He or she at times experiences difficulties on one side or the other. There is no good reason to hide or ignore the resulting tensions, nor to make fuss about them. We shall draw some consequences of this role of advisers at the end of the next and last section of this book.

4.6.2. *Towards a closer view of the interaction between models and policies*

What was often called the Frisch–Tinbergen theory of economic policy was exposed here in Subsections 3.6.2 and 3.6.3, when it naturally fit into the development of our study of structural macroeconometric models. Indeed, the three basic elements were a selection of variables, with in particular instruments and target variables, a model of the economy, an objective function. The choice of a policy was formalized as a mathematical optimization problem: the values of the instruments had to be chosen in order to maximize the objective function under the constraints imposed by the model.

Such a formalization provides a correct account of the underlying logic of policy interventions but not a close description of the actual decision process. Under (ii) in Subsection 3.6.3 we already recognized that, first, the objective function of deciders is not made explicit by them, second, analysts or advisers have to infer it from past choices or declarations of those deciders, and third, such inference is real as a mental operation but was seldom made explicit as a formal procedure. About the actual uses of a model in the decision process our examination was hardly less remote. We discussed in Section 3.5 solutions obtained from the model for alternative sets of values of the instruments, thanks to mathematical simulations in most cases. Again the vision conveyed by such an examination, although fundamentally adequate, ignores features of the interaction between the analysts in charge of the model and advisers or even deciders.

A recent book, edited by F. den Butter and M. Morgan, may help the reader to know these features and their institutional context[271]. Here we cannot really go into the study of institutional arrangements: this would lead us outside our domain and require references to national specificities. But we shall draw from the book three important ideas which seem to apply almost everywhere.

(i) *A non-negligible dose of eclecticism now surrounds the use of a model for policy purpose.*

This was not always so. Often in the past the Frisch–Tinbergen methodology was uncritically applied, the model being a black box for the users. But this could not last long because policy advisers and deciders wanted to understand why a simulation was announcing that such and such results would follow from a particular policy package. Everywhere an interchange had to occur in order to explain the results. The criticisms raised against structural macroeconometric models, to which we referred in Section 3.8, made users of these models more circumspect and more prone to request explanations.

The practice, in treasuries and central banks, to work with one main model was not often discontinued, but the teams in charge of operating the model and interpreting its results more and more often now use in parallel other macroeconomic tools. These may be one, two or several smaller-size models each one geared to a specific purpose: a better representation of long-run trends, a more detailed analysis of the wage-price inflation spiral taken separately, an investigation into the determinants of the exchange rate, and so on. Some of these additional models may be just "calibrated" instead of being econometrically estimated from the aggregate time-series base. Often also simple VARs or "structural VARs" are computed and analysed. Moreover, in some countries, public and private owners of macroeconometric models have the habit to bring together and compare results delivered by their respective models. The organization of such meetings may even be instituted as a formal part of the public decision-making process.

For instance, in the den Butter–Morgan book, P. Duguay and D. Longworth report on "Macroeconomic models and policy making at the Bank of Canada". They explain how econometric model building at the bank began in the mid-1960s, and how since then policy, model use and model development were all related. In the late 1970s the staff recognized that the main model (a second version of the one initially built) had not kept up with developments in the

[271] F. den Butter and M. Morgan (eds.), *Empirical Models and Policy Making; Interaction and Institutions* (Routledge, London, 2000). See in particular the long synthesis offered by the editors, also published as: What makes the models-policy interaction successful?, *Economic Modelling* (July 1998).

Bank's thinking and in monetary economics more generally. Simpler calibrated simulation models, the dynamics of which were easier to figure out, began to be built for use in policy analysis. In the early 1980s a "small annual model" was set up, after serious theoretical discussions and definition of its steady-state characteristics, for use in medium-term policy simulations which could help address questions that preoccupied the policy makers. In the middle 1980s the main vehicle for comparison of alternative nominal targets and monetary policy rules was a quarterly calibrated model with four key equations plus a policy rule and identities; expectations were assumed adaptive in the short run and model-consistent in the long run, this in order to rule out the notion of costless disinflation implied by the pure rational expectations hypothesis. After a survey of the views held by users of the main econometric model in the late 1988, it was decided to build a new model which became the main tool for projection in 1993. The parameters of the model were calibrated so as to fit with what was learned from a number of empirical sources, including vector autoregressions. Satellite models now serve for disaggregations. Smaller models approximating the behaviour of the main model are used for stochastic simulations. All these developments are seen as resulting from the interaction between policy makers and the analysts working at the Bank.

Eclecticism is particularly deliberate at the Bank of England in a team the approach of which is announced to be "multi-model"[272]. Five categories of models are distinguished, and ranked here from the most to the least theory-loaded: small analytical models, inspired by the practices of the RBC movement; stylized macro models, for instance one inspired by the "Dornbusch overshooting model" (five equations determining real output, the quantity of money, the level of prices, the interest rate and the exchange rate); macroeconometric forecasting models, i.e. structural models of the traditional type; simple output-gap models, actually two equations relating inflation to measures of the degree of capacity utilization; VAR models of inflation. The macroeconometric forecasting models provide the central tool, but as a flexible instrument with no unique model; the models are much smaller than other macroeconometric models of the UK, disaggregations being handled in series of submodels; they are there to provide a benchmark for average responses to average shocks and to map the analysis from other models into inflation projection.

(ii) *Usually, however, a central macroeconometric model serves as framing the decision process*. In the den Butter–Morgan book, R. Smith wants to explain why macroeconometric models have survived and prospered, despite

[272] J. Whitley, Economic models and policy-making, *Bank of England Quarterly Bulletin*, 1997, 163–173.

their deficiencies, the criticisms raised against them in some academic quarters and the scepticism shown by a number of users.

He draws a parallel with the management literature where it appears that the formal systems taught in textbooks are useful abstractions rather than accurate descriptions: strategies actually emerge from a set of rather informal, interactive relationships as the agents respond to events; formal systems are ritualistic, but nevertheless provide a framework for developing the emergent strategy. Hence the title of his article: "Emergent policy-making with macroeconometric models". In order to make his point Smith studied published statements of modelling teams, mostly memoranda submitted by advisers to the Treasury in the UK. He shows that "macroeconomic policy emerges in many different ways from a variety of sources, not least the pressure of events. The announced policy is the product of an organisation... a model can help... in storing and sharing information, provide structure and discipline to the policy evaluation process, provide a focus for dialogue about particular issues under dispute, ensure coherence and consistency and aid the learning process... Policy making is inherently quantitative and needs numbers. A model... is also a flexible way, allowing the input of judgement and the exploration of alternatives". In short: "The model provides information that helps answer the question, it does not provide answers".

Before drawing the conclusions so quoted, R. Smith also writes: "While most organisations will, quite rightly, use different models for different types of question, there is almost certainly going to be a role for models that look like traditional econometric models, because their outputs are tailored to the policy process in a way that most alternatives, e.g. ARIMA models, are not".

It would be easy to collect testimonies of economic advisers conveying similar messages. In the den Butter–Morgan book we find other supports to about the same conclusions. In particular, under the title "US monetary policy and econometric modelling: tales from the FOMC transcripts 1984–1990", H. Edison and J. Marquez conclude: "What emerges from this study is that models bring to policymakers internal consistency and quantifications. In addition, policy makers pose interesting questions to modelbuilders and provide suggestions for enhancing the role of models for policy discussions. We conclude that models are used to shape judgements and are not mechanistic providers of policy constraints". This leads us to our next point also.

(iii) *Models are revised in response to needs or deficiencies revealed by the interaction with policy making*. In the den Butter–Morgan book this aspect of the interaction is closely studied by P. van Els writing under the title "Policy making and model development: the case of the Nederlandsche Bank model MORKMON" (read in English: MOnetary-Real Quarterly MOdel for the Netherlands). As the author notes, limiting attention to one model only (actually the main model of the central bank) means overlooking the fact that policy making influences also the choice of the set of models used in the institution, a comment about which we shall say more later. But the discussion of the article now examined is so more focused.

In order to identify the linkages between policy and model development, van Els has to take account of other forces which influenced this development, particularly trends in macroeconomic and econometric theory, and changes in the general attitude among academics towards the design and use of the model. We recognized this influence in Section 3.8 here, when we listed five main lines along which the representation of phenomena in structural macroeconometric models was revised during the last two decades. Two of them were clearly responding to academic criticism (attention to the long-run implications of the wage-price dynamics, modelling of expectations formation). Academic research and policy making were both responsible for two other lines of revision (modelling of "the supply side" and the labour market). Only for the last one (a better representation of policy instruments and their initial impacts) was the demand of policy the prime mover.

After his survey of non-policy influences on the development of MORKMON, van Els begins with explicit linkages between policy and this development. He then presents examples concerning monetary or fiscal policy. Let us quote three of them. Since the mid 1980s the Dutch monetary policy was almost entirely focused on the exchange rate with the Deutsch mark as an intermediate target; this required the reaction function of the domestic short-term interest rate to involve the spot premium of the guilder against the DM and the German short-term interest rate. In the late 1980s a model version was developed for the analysis of a new anticipated regulation of commercial bank reserves at the central bank (replacement of direct credit ceilings by penalties on excess growth of money-creating operations above a certain threshold percentage); the effectiveness of the new regulation was found to depend crucially on its impact on long-term interest rates. The fiscal part of the model had to be developed so as to single out in particular employers' contributions to social security, the rate of which was frequently used as a policy instrument to control the evolution of labour costs.

Other linkages, called implicit linkages in the article, also play important parts. Model revisions are sometimes triggered by views expressed by policy makers. For instance, monetary authorities frequently emphasize macroeconomic and monetary stability as a prerequisite for a favourable economic performance. The model was used as a research tool to investigate the empirical validity of this policy presumption. The possible role of stability indicators was explored. Confidence, measured from changes in unemployment and public deficit, has been found to be a significant determinant of private consumption, and exchange rate volatility to significantly affect exports.

Progressive improvements in the quality of the model or in practices about its use take account of policy concerns in a more diffused way, which is difficult to disentangle from non-policy influences. For instance, the Lucas' critique had a direct impact on the way the analysts revised MORKMON. But, in addition, it affected the way of using the model for policy analysis. Considering that "the major merit of the Lucas critique [was to warn] against a too mechanical use of the model, in particular when exploring more radical policy changes, it has been common practice [in such cases] to base policy advice not on one simulation exclusively but to… conduct sensitivity analyses with respect to crucial parameter values; to consult different model-versions or even models from other institutions".

Finally the policy purpose is part of the explanation why the central bank developed an own model rather than hiring the services of outside model owners. "Model building activities generate positive external effects on other research activities. Databases have to be constructed which can be used for other purposes… Building an own model brings with it full control over its operation. This enhances flexibility and decisiveness in model use, which are essential factors for a successful contribution of models to the policy design process."

4.6.3. *Lessons to be learned from the model-policy interaction*

We are now in a position to resume the assessments which were the subject of Section 3.8 where we ended Part 3 devoted to structural econometric models. We shall begin with a brief survey of a debate which took place in the Netherlands, the country where those models were born. We shall later comment on the distinction between the two main purposes served by structural models: forecasting and policy analysis.

(i) The remark that macroeconomic analysts have lessons to learn from experience of the model-policy interaction says hardly more than was already implied in the preceding subsection or even in Subsection 3.8.2. But the den Butter–Morgan book highlights the remark in two interesting ways: how lessons might be learned from competition between models aimed to serve the same purpose (rather than from the complementarity that may appear implicit in the eclectic multi-model approach), how a tension might appear in the development of a model because different users of this model may have different priorities in their ranking of alternative developments.

The first part of the book concerns "*The Tinbergen tradition*" *in the Netherlands*. After an informed address of the Minister of Finance G. Zalm, a former director of the Central Planning Bureau (CPB), the article by J. Donders and J. Graafland surveys the experience of this CPB, which was created immediately after the second world war with Jan Tinbergen as its first director, and since then skillfully followed the tradition. The article by P. van Bergeijk and J. van Sinderen describes how a number of macroeconomic models were developed since 1980 outside the CPB, confronts their results to those produced by the CPB models and argues that the latter were not suited for the analysis of some important structural policy issues. Considering the point we now want to study, we shall pay particular attention to this thesis of what we shall call here the vBvS article.

Readers should probably remember now not only the Tinbergen tradition, but also the "Dutch disease" of the early 1980s. The phrase characterizes the situation when the unemployment rate peaked at 11.9 per cent in 1982. Incomes earned from the North sea oil and gas fields since the late 1970s were said to be responsible for an overvaluation of the guilder and a lax behaviour of employers in wage negotiations. The cure came from a complete change in people expectations and attitudes, which was manifest with the "Wassenaar agreement" reached in the autumn of 1982 by the central organizations of employees and employers, then threatened by a policy package of the newly appointed government which considered a private-sector wage and price freeze. The agreement traded automatic cost-of-living indexation for labour time reduction. It was followed by years of substantial wage moderation, with government cuts in public sector wages, the legal minimum wage and related social security benefits. At the end of 1999 the Dutch unemployment rate was down to about 3 per cent.

According to Donders and Graafland the Dutch disease was announced to officials by CPB in the summer of 1975: "not only would strong wage increases shorten the economic life time of capital goods, but they would also reduce profits and, hence, the creation of new jobs by investments". Development of models at CPB during the second half of the 1970s was motivated by this diagnosis. It evoked in 1977–78 a vivid discussion among academics and government specialists. Thus CPB analyses showed the need for a change of economic policy already in an early stage, long before a substantial policy change occurred.

The argument of the vBvS article concerns neither that period nor wage policy, but rather the kind of structural reform that was discussed later and for the study of which there was indeed considerable demand from policy makers and the public. More precisely the vBvS article refers to three such structural reforms: the reduction of government expenditures publicly discussed in the 1980s, competition policy which was on the political agenda in the 1990s, finally a regulatory energy tax. The second case, which is presented in just two paragraphs, has ambiguous significance, because assessments of the macroeconomic impacts of such reforms as liberalization of shops open hours or deregulation of telecommunication have not yet proved to be reliable: the statement by CPB that it did not know how to quantify the effects of such changes in the legal framework may be taken as a sign of honesty rather than failure to keep up with macroeconomic knowledge.

For the reform of public finances the vBvS article takes the example of the effects predicted by various Dutch models after an hypothetical reduction of direct taxes financed by a simultaneous decrease in public spending. The CPB models announce a (slight) decrease in employment whereas some other models predict an increase. In most cases the impact on employment is small, but three models predict a significant positive increase. As far as this author understands, the distinctive feature of the latter models is to embody a strong negative impact of public expenditure on investment; this followed the opinion, often heard in the early 1980s, that the excessive burden of public spending was structural and crowded out private expenditure, especially private investment, an argument which was formalized and empirically tested by A. Knoester[273]. At the time this was the opinion of the cabinet; Donders and Graafland describe the critics then voiced against the CPB and how model development somewhat changed the CPB assessments, but not much (a reduction in the fiscal deficit would imply a short-term pain in order to realize a long-term gain). Readers of this book may remember Section 7.1 of Chapter 7, in which we discussed the challenging expansions that followed some severe fiscal contractions. Our report then gave considerable weight to the restoration of confidence in the country and to the resulting change in private expectations. (Remember: according to R. Solow quoted at the beginning of our Section 4.2, those "things" may be fleeting and so put models out of adequacy.)

P. van Bergeijk and J. van Sinderen make interesting comments after their report of model assessments about the macroeconomic impact of a regulatory energy tax, the third structural reform they examine. They give the output response to a tax shift from labour to energy, as predicted by six models. The results vary widely, although none of the models announces a positive response. The variation will not be a surprise to those who know how difficult it is to build a reliable theory of the incidence of various taxes in an economy experiencing

[273] A. Knoester, Stagnation and the inverted Haavelmo effect: some international evidence, *De Economist* **131** (1983) 548–584.

market disequilibria. But the authors add: "We consider the pluriformity of analyses as an asset for the Dutch policy process as it makes it possible to provide for a cost efficient assessment of policy alternatives and their robustness. Pluriformity also diffuses the knowledge about macro modelling, generating expertise outside the CPB. Also in this sense the increased competition in the macromodelling industry has added value to the policy process". This quote brings in a new argument with which the author of this book fully agrees: it no longer concerns the predictive value of various models, but rather the public visibility of what macroeconometricians can achieve.

(ii) Different models and even a particular model may serve different purposes. The idea was present earlier at various places in the preceding subsection, but it deserves a fuller attention and a methodological digression stimulated by an article contained in the den Butter–Morgan book: "*Conflicts between macroeconomic forecasting and policy analysis*", by S. Wren-Lewis.

We read in the abstract: "Econometric macroeconomic models tend to be used by policy-makers for two purposes: forecasting and policy simulation... This paper argues that, as models developed, important conflicts emerged between forecasting and policy analysis. Pressures and incentives within the policy making process led to an emphasis on the forecasting role, such that the ability of these models to provide policy advice was compromised... The solution to these problems is to develop different models for forecasting and policy simulation". The thesis looks interesting, but the argument of the article reveals to be somewhat misdirected. The author argues that there was an initial misconception in the research program of the macroeconometric movement: "the idea that theory and econometrics would combine in a harmonious way", but, Wren-Lewis writes, "it was quite possible to produce reasonable forecasts with little or no reference to economic theory, and it was quite possible to give policy advice with no reference (in the form of conventional econometric estimation) to the data".

With little exaggeration we may say that the argument seems to claim that data without theory are good for forecasting and theory without data is good for predicting the effects of policies. A reaction immediately comes to mind: what does the author mean by theory? We have two clues for the answer: "... existing theories [have] an alternative source of self justification besides the data, which [is] the rationality principle", and, in a different article which aims at linking econometric models with theory: "most current macroeconometric models have inconsistent or old-fashioned theoretical foundations"[274].

[274] Page 543 in S. Wren-Lewis, J. Darby, J. Ireland and O. Ricchi, The macroeconomic effects of fiscal policy: linking an econometric model with theory, *The Economic Journal* (May 1996).

I shall stop again on this argument, noting in passing that for the author old-fashioned means bad, and that indeed the second article identifies theory with what is proposed in a *recent* article. Rather than expanding what I wrote on the subject of microeconomic foundations in Section 3.8, I shall argue that the criteria for evaluating the applications of macroeconomic theories are fundamentally the same as those for evaluating macroeconometric models. Following H. Pesaran and R. Smith[275], we want macroeconometric models to be, first, *relevant* to the purpose for which they were built, and so we would like theories to be to the problems they claim to tackle in applications. We want models to be, second, *consistent* with what we believe to know about the phenomena, and theories are of course meant to be expressions of that knowledge, which contains a lot in particular on how to realistically apply "the rationality principle". Since econometric models are used for quantitative assessments, we want them to provide, third, *adequate* quantitative representations of the phenomenon of interest, and this in all the various respects addressed by econometric inference procedures, i.e. selection among models that are equally relevant and consistent, estimation of parameters and test of particular hypotheses, which would make the model more specific, and often also more transparent or easy to use. Similarly, if theory is meant to be applied to policy evaluation, it has to become somehow quantitative, which may be obtained in various ways all relying on the same fundamental principles, as we saw in Section 4.5.

This being said, the thesis put forward by Wren-Lewis concerns how to gauge the respective mixes of relevance, consistency and adequacy for a given macroeconomic instrument applied to two different purposes, forecasting and policy analysis. He claims: macroeconometric models are good for forecasting, bad for policy analysis; macroeconomic theory is good for policy analysis, bad for forecasting. It seems to me that he would be on a better ground if he would speak of a different (and old) distinction between purposes: the short term and the medium term. Moreover the distinction would capture a good deal of what he has in mind.

It is true that pure data analyses, of the VAR type in particular, do not perform badly for the very short run, at least in normal times. We discussed at length the forecasting performance of macroeconometric models in Section 3.4. We then concluded that it is definitely superior to that of VARs at the horizon of two years, say. In particular these models are designed and used in such a way as to naturally embody the different kinds of shocks that may occur and be quickly characterized. Hence it was not surprising to record in

[275] H. Pesaran and R. Smith, Evaluation of macroeconometric models, *Economic Modelling* (April 1985).

Section 3.5 that for short-term policy analysis also the performance of structural models is good, relatively to what alternative macroeconomic tools can offer.

But in the same Section 3.5 we saw that medium-term projections as well as medium-term simulations given by structural models were much less reliable, in particular because different models led to different results. We then saw how such differences could be traced back to differences in the specifications of the models, as well as in the estimates of some parameters. In Subsection 3.8.3 we commented at some length on the difficulty of medium-term forecasts and simulations[276]. But we concluded that the difficulty was inescapable.

Was this last conclusion erroneous? Can "theory" give us about the medium run reliable assessments? If so, will that be without the assistance of a good deal of econometric work on real data? Or rather, why could not a good macroeconometrician build a structural model having medium-run forecasts and policy simulations as its purpose? Could not this macroeconometrician match for the purpose the appropriate theory with the appropriate data? These questions are put here as challenges to the profession, perhaps in particular to Simon Wren-Lewis. Remember also that, right after commenting on the inescapable difficulty of medium-term assessments, I advised macroeconometricians to avoid overconfidence and to be open to the possible application of new ways of modelling phenomena.

Just in passing let us also note that, in the process of elaborating a theory of economic growth, business cycles, inflation or unemployment, we are often led to raise side questions of a purely logical nature. Quite naturally research addresses such question and, in case of success, articles are published which provide answers to a given side question. But such interesting results, internal to the development of a theory, are seldom directly useful for applications of the theory.

Let us now follow the interesting approach taken by Wren-Lewis and consider past trends in the demands addressed to the market for the services of applied macroeconomists. Is there now a more serious conflict than in the past between the demands of policy makers and those of private users? Probably so, but the conflict should not be overemphasized. Moreover, the best way to cope with it is not to rule out the present macroeconometric models from the policy simulation business.

[276] Let us note in passing that we then also recognized that current macroeconometric models were not well designed for analyses of structural reforms, which were at issue a moment ago in our report of the Dutch debate. Analysis of such reforms has indeed more to draw, in the first place, from microeconomic than from macroeconomic theory.

In the older days most policy makers requested from macroeconomists hardly more than short-term analyses: the macroeconomy had to be controlled, but that could be done at fairly short notice; being alert and properly handling the instruments would suffice. Attitudes certainly changed for various reasons: the realization that lags in the effects of instrument manoeuvre were longer than had been expected; hence a policy of fine tuning could turn out to be destabilizing; such appeared to be the case during the 1960s and 1970s in some countries like the UK; other macroeconomic problems than those coming from changes in aggregate demand were identified, for instance, with the Dutch disease or the need to durably cure inflation, and their treatment was a longer-term affair. All this meant that devoted policy-makers had to worry more about the medium run than the short run.

However, it does not follow that demands to owners of macroeconometric models no longer came from government but rather only from a profit-seeking private sector. Indeed short-term forecasts remain useful for government managers, just as they are for business managers, for instance in financing current operations. On the other hand, there is also an important private demand for medium-term forecasts, which are needed for decisions about all kinds of investments.

Moreover we saw, a moment ago about the Dutch case and earlier when discussing in Section 3.8 the model developments of the last two decades, that macroeconometric models can be made suitable to a significant range of medium-term policy analyses. On the other hand, we have to see better the "different models" which, according to Wren-Lewis in his abstract and text, ought to take over policy simulations: what will these models exactly be? How well will they perform?

4.7. Macroeconomic public finance strategies

In the last section of a long book the author may be tempted to look back and to recognize the imperfections of his work. But no attempt at listing them will be made here. More important is the question to know whether the teacher conveyed to his students a correct idea of the field he announced to cover, a field of knowledge which is full of gaps calling for research, but in which some solid landmarks also exist. Learned readers will certainly find cases in which they would like to dispute the description given of such landmarks. New-comers on their part may want to go more deeply into the subject and to investigate some of the points which were mentioned. There is indeed still much to do for the diffusion and progress of macroeconomics.

Cases certainly also remain of important landmarks which were ignored. Given that here is the last opportunity for complementing the text, the author will still add a few bits concerning points that might have deserved fuller consideration. The common feature of these bits is to deal with macroeconomic policies viewed in a dynamic context. We may speak of macroeconomic policy strategies, or more simply of macroeconomic strategies. Since monetary policies were discussed at length, particularly in Chapter 8, we shall now rather focus on fiscal policies, i.e. on the long-run management of public finances and on their short-run regulation. This will provide the opportunity of touching on political economy arguments, which quickly come to mind in the discussion and have recently been the subject of substantial research. We shall end up with a few words about the ethics of policy advising.

4.7.1. *Arithmetic of the public debt*

Our entry into the discussion of fiscal strategies may be borrowed from B. Trehan and C. Walsh[277] who make the two following common-sense remarks: "The intertemporal budget constraint imposes restrictions only on the long-run relationship between expenditures and revenues so that almost any short-run deficit path is consistent with a budget balanced in present value terms... But expenditures (appropriately defined) cannot drift too far away from revenues (appropriately defined)". On the one hand, governments can always devote huge amounts of borrowed resources to finance perceived emergencies. On the other hand, most governments have to face the large cost of their public debt and are permanently fighting to contain it by cutting expenditures or raising taxes. What should a wise benevolent government do? What are actual governments doing?

(i) In order to correctly pose and answer such questions, we must first look at the accounting equations ruling the path of the public debt and at the long-run intertemporal budget constraint. These can be formalized in various ways, depending on the definition of such aggregates as "expenditures", "revenues", "public debt". We shall consider a particular formalization and later see how it relates to those introduced earlier in the book, particularly in the first sections of Chapter 1.

[277] B. Trehan and C. Walsh, Common trends, the government's budget constraint, and revenue smoothing, *Journal of Economic Dynamics and Control* (June–September 1988).

For a start let us focus on the public debt, at the beginning of period t and denote it as B_t. The common measure of the size of this debt is b_t its ratio to GDP in period t:

$$b_t = \frac{B_t}{Y_t}, \quad Y_t = p_t y_t. \tag{330}$$

During period t government has to pay interest $r_t B_t$ to its creditors, r_t being the nominal interest rate, actually an average over interest rates applying to various types of debts, some of them being held by creditors in the form of long-term bonds, others in short-term Treasury bills. The net change in the public debt may be written as:

$$B_{t+1} - B_t = r_t B_t - S_t, \tag{331}$$

where S_t may be called for the time being the surplus on current operations, a surplus available at the end of period t (we shall say more about it in a moment). The corresponding equation for the size of the debt is:

$$(1 + g_{Yt})b_{t+1} = (1 + r_t)b_t - s_t \tag{332}$$

with

$$1 + g_{Yt} = \frac{Y_{t+1}}{Y_t}, \qquad s_t = \frac{S_t}{Y_t}. \tag{333}$$

From the recurrence equation (332) we can compute the size of the debt b_T starting from b_1 in the initial period with the surpluses $s_1, s_2, \ldots, s_{T-1}$ in the meantime. This is easily done if we write:

$$a_t = \frac{1 + g_{Yt}}{1 + r_t}, \qquad \gamma_t = \prod_{\theta=1}^{t-1} a_\theta, \qquad \hat{s}_t = \frac{s_t}{1 + r_t}. \tag{334}$$

We note that $\hat{s}_t$ is the present value at the beginning of period t of the size of the surplus, which will be available at the end. We may write (332) as:

$$a_t b_{t+1} = b_t - \hat{s}_t$$

or better

$$\gamma_{t+1} b_{t+1} = \gamma_t b_t - \gamma_t \hat{s}_t, \tag{335}$$

which, with $\gamma_1 = 1$, directly gives:

$$\gamma_T b_T = b_1 - \sum_{\theta=1}^{T-1} \gamma_\theta \hat{s}_\theta. \tag{336}$$

The sum in the right hand side is similar to a present value of the flow $\hat{s}_1, \hat{s}_2, \ldots, \hat{s}_{T-1}$, except that the growth rates of nominal GDP appear in the numerator of a_t.

Hence, we immediately see that there are two factors acting on the evolution of the size of the debt: the sum in the right-hand side, and the value of γ_T. Other things being equal, b_T is a decreasing function of γ_T. Usually we expect the numbers a_t, hence γ_T, to be smaller than 1, because we expect the nominal interest rate to be larger than the rate of growth of nominal GDP. If such is the case, a decrease in the size of the public debt cannot occur unless the pseudo-present-value of the flow of $\hat{s}_t$ is positive and sufficiently large. However, we experienced periods during which the rate of the interest to be paid on the public debt was smaller than the rate of growth, often because of a dominant accelerating inflation. In those times few governments were concerned by their public debt. The situation radically changed in the 1980s because of the world-wide rise in interest rates and the policies of disinflation. Then in a number of countries interest to be paid on the public debt was found excessively high and the size of the debt increased at a speed that appeared unsustainable, except for a very short period.

From (336) we also derive a natural expression of *the intertemporal budget constraint* that managers of the public debt ought to take into account in period 1, namely:

$$\sum_{\theta=1}^{\infty} \gamma_\theta \hat{s}_\theta = b_1. \tag{337}$$

Indeed, if most numbers a_θ in the computation of γ_T according to (334) are smaller than 1, γ_T tends to zero. So b_T would increase beyond any limit if the right-hand side in (336) would remain positive without tending to zero.

Note that in (337) the left-hand member has to be understood in such a way as to recognize the uncertainty of expectations of future periods: in other words, except perhaps for $\theta = 1$, the product $\gamma_\theta \hat{s}_\theta$ is a random variable, which is commonly replaced by its mathematical expectation, conditional on information in period 1. In the literature the assumption of constant expectations about the real growth rate and the real interest rate is often made, leading to $\gamma_t = a^{t-1}$.

Before discussing the meaning of the surplus appearing in Equations (331), (336) and (337), let us briefly refer back to Chapter 1. Although we then made explicit the relations between assets and operations, there are various reasons why we did not write this equation. The main reason was a broader definition of government, which then even included all financial institutions. But the accounting framework of Chapter 1 also draws our attention to two complications.

First, Equation (331) identifies the net issue of government bonds with the change in the value of the public debt. This ignores the possibility of capital gains or losses on the value of the outstanding debt (compare with Equation (5) in Chapter 1). In other words it implicitly assumes that B_t is the par value of the debt, not its market value, whereas when the central bank buys government bonds by open market operations it pays them the market value. Is this choice of the par value a real problem in our discussion here? The question amounts to knowing whether speculative bubbles on the price of government bonds may play a significant role. We discussed the presence of speculative bubbles more generally in Section 1.8 of Chapter 8. Theoretical investigations showed the logical possibility of a persistent bubble on public bond prices. But available empirical tests suggest that this possibility does not materialize[278].

Second, the accounting framework of Chapter 1 shows that, in contrast, we cannot neglect the issue of money by government, even with the narrow contour of government which is meaningful in our present discussion. Indeed, government finances its expenditures from three sources rather than two: not only taxes and increase in its debt, but also part of the increase in the money stock: issue of currency notes and coins *plus* increase in reserves of commercial banks at the central bank. The part of the money stock that is so involved is commonly called "base money" or "high-powered money". We denote it M_{gt}.

The corresponding extra resource, which is often called "seigniorage", has clearly to be taken into account, since it was quite substantial in some cases (like 3 per cent of GDP in a number of European countries during the stagflation that occurred around 1980[279]). We shall write here:

$$\sigma_t = \frac{M_{g,t+1} - M_{gt}}{p_t y_t}. \tag{338}$$

The surplus S_t then has to be equal to seigniorage *plus* taxes $(\tau_t p_t y_t)$ *minus* public expenditures, written here as $x_t p_t y_t$. Hence:

$$s_t = \tau_t + \sigma_t - x_t. \tag{339}$$

So defined the surplus S_t does not correspond exactly to the government budget surplus, or to the opposite of its deficit, as usually recorded. In our notation that public deficit would be equal to $(x_t + r_t b_t - \tau_t) p_t y_t$: it would include

278 See J. Hamilton and M. Flavin, On the limitations of government borrowing: a framework for empirical testing, *American Economic Review* (September 1986).

279 See Table 16 in N. Roubini and J. Sachs, Government spending and budget deficits in the industrial countries, *Economic Policy* (April 1989).

interest of the debt as expenditure and exclude seigniorage from revenue. Another familiar concept in public finance is the *primary surplus*, which excludes interest of the debt from expenditures, hence in our notation $(\tau_t - x_t) p_t y_t$.

Finally, the reader notes that, when in this section we are speaking of public expenditures (exclusive of interest of the debt), we are referring not only to public consumption but also to transfer payments to private agents, i.e. to the sum of the two aggregates X_t and S_t of the accounting framework of Chapter 7. Correlatively, taxes are meant not to be measured net of transfer payments (such a net concept was used in the accounting equations of Chapter 1).

4.7.2. *Sustainability of the public debt*

It frequently occurred in the past, particularly in the last two decades, that citizens and business analysts of a country expressed worries about excessive public deficits and increasing public debt of their country. The concern had grown out from the feeling that public finances were engaged in a process which could not be long sustained. Can we say that such fears are unwarranted because the intertemporal budget constraint leaves ample room for repaying current debts later? It would be too easy, because other considerations, not brought out by the foregoing arithmetic, cannot be neglected. Although soft, the concept of sustainability of public finance makes sense because of two main reasons: beyond a point the marginal cost of servicing the debt is likely to start increasing, and the freedom of government to cope with depressive shocks is likely to be reduced.

At low levels of the public debt, government bonds are quite safe assets except for the risk of inflation. But this risk sharply increases when the debt is high or seems to grow beyond control. Lenders then start requesting risk-premia, especially foreign lenders, so that the rate of interest on the debt increases, a first effect which magnifies the public finance problem. Along this cumulative trend the credibility of government as an honest borrower also decreases.

If public opinion feels that the government deficit is excessive, any depressive shock, which will naturally trigger “automatic stabilizers”, will also worsen the deficit and aggravate the public concern. Ministers will then find their hands are too tied to allow any expansionary fiscal move. They will rather think of cutting expenditures or raising taxes in order to save their fiscal credibility.

Clearly these two kinds of effect are mutually reinforcing.

Such considerations motivated the search for empirical indicators which would signal when and where the risk of unsustainability emerged. Two approaches were then used. The first one looked at the path of the debt-GDP ratio b_t, or at the path of a similarly relevant statistics. A problem would be identified as having emerged in a sample period (t_0, t_1) if the hypothesis had to be rejected, at usual levels of significance, according to which the path of b_t during the period would have been just the realization of a sustainable random process. In the tests performed by G. Corsetti and N. Roubini[280] for the period 1971 to 1989, the linear stochastic process assumed to have generated the series was considered to be sustainable if either it was stationary or its first difference was stationary with a zero mean. (In the maintained hypothesis b_t was the sum of four terms: a constant, a deterministic linear trend, a term proportional to b_{t-1} and a stationary linear process with zero mean.) The test rejected sustainability of the path followed by b_t, during the two decades, for four Western European countries, the two most manifest cases being Italy, where b_t had increased rather regularly from 0.43 to 0.96, and Belgium, with an increase from 0.63 to 1.26. We may wonder whether such a mechanical test on the time series of b_t adds anything significant to direct inspection of the series, to which indeed the authors pay great attention (they even recognize that their test failed to "detect the structural change in Belgium's fiscal policy that has occurred in the last few years").

Another approach focused on the concept of a fiscal "primary gap" at date t, defined as the difference between the actual primary surplus (as usually measured at t and as a ratio to GDP) and a constant level of the primary-surplus ratio which would lead, at a conventional future horizon, to the same level b_t of the debt-GDP ratio as at date t. The measure of this primary gap clearly depends not only on the choice of the horizon but also on assumptions about the future evolution of $r_{t+\theta}$, $g_{Y,t+\theta}$ and $\sigma_{t+\theta}$: this follows from Equations (332) and (339). In particular a "myopic permanent primary gap" is defined if the horizon is infinitely far in the future and given constant (independent of θ) values are assumed for $r_{t+\theta}$, $g_{Y,t+\theta}$ and $\sigma_{t+\theta}$.

Calculations were made for such a myopic permanent primary gap in the year 1992 assuming a future real growth rate of 3 per cent, a future real interest rate of 5 per cent (both annually) and a zero seigniorage, this for the twelve countries of the then European Union[281]. At that time most of those countries had a positive primary surplus, even a substantial one in five of them. This explains why a positive gap, i.e. the need to raise further the primary surplus for long-run stabilization of b_t, appeared for only five countries. For instance,

[280] G. Corsetti and N. Roubini, The design of optimal fiscal rules for Europe after 1992, in: F. Giavazzi and F. Torres (eds.), *Adjustment and Growth in the European Monetary Union* (Cambridge University Press, 1993).

[281] See W. Buiter, G. Corsetti and N. Roubini, Excessive deficits: sense and nonsense in the Treaty of Maastricht, *Economic Policy* (April 1993).

in Italy, where the primary surplus ratio was equal to 0.7 per cent, a gap of a further required increase amounting to 1.2 per cent remained. But in Belgium the primary surplus ratio, equal to 5.4 per cent, well exceeded the permanent required primary surplus ratio, evaluated at 2.5 per cent; the primary gap was so equal to -2.9 per cent: this is consistent with the fact that, after peaking at 1.35 in the depression year 1993, the Belgium debt-GDP ratio decreased (1.18 in 1998).

This examination of sustainability indicators for Western European countries gives us the opportunity to note that large differences across countries were observed. For instance, whereas substantial increases in b_t between 1971 and 1989 were recorded in almost all countries, the levels reached in 1989 differed (with 0.34 in France against 1.26 in Belgium) and the UK ratio had decreased during the period from 0.76 to 0.38.

These differences are a sign that, notwithstanding the arguments given at the beginning of this subsection, sustainability and credibility considerations do not suffice to remove the feasibility of large intertemporal shifts in the management of public finances.

4.7.3. *Tax smoothing*

Historical data show that, at least in cases of important wars when public expenditures for defense increased to exceptional levels, government borrowed large amount which were later repaid on a number of years thanks to the revenues generated by a relatively moderate increase in the tax rate. The tax-smoothing hypothesis states that more generally, in order to finance an irregular flow of expenditures, governments choose to smooth the evolution of the tax rate, within the limits permitted by the intertemporal budget constraint. This looks like a natural hypothesis to make. Why so? Is it confirmed by the data?

(i) Abstracting from political considerations, we may find economic arguments for tax smoothing in preference to choosing a path of the tax rate that would closely track the path of the expenditures to GDP ratio. Neglecting seigniorage which is small except in times of inflation, we may as well compare a policy of tax smoothing with one imposing a balanced budget every year ($\tau_t = x_t + r_t b_t$). Then the business cycle is, besides wars, a second major case in which the tax-smoothing policy deviates from the balanced-budget policy. Indeed, inspection of time series also supports the idea that the contracyclical fluctuations in x_t (generated by the GDP cycle) or in $x_t + r_t b_t$ are not reflected in the path of the tax rate.

Aggregate demand analysis suffices to explain this second observation. Systematically raising tax rates during recessions would make recessions worse, since it would amount to cancel the effect of automatic stabilizers. To this extent at least, the conclusions of Keynesian teaching about fiscal policy is almost unanimously accepted among policy makers. Demand analysis should not be neglected in a close study of business conditions during war times. Indeed, part of the cost of the second world war in the US was financed by revenue resulting from a higher degree of activity. But the part financed by increases in the public debt was clearly more important. We must therefore look at another kind of argument in order to explain why tax rates were not then raised in line with the increase in public spending.

(ii) An argument was given by R. Barro, suggesting that, acting as a social planner in a neoclassical competitive economy, the government should choose to smooth taxes when it needs to finance an irregular flow of public expenditures[282].

Adopting a general equilibrium approach, we might view the problem as belonging to the theory of optimal growth. The social planner, while aiming at maximizing the utility over time of a representative individual, would control only "the tax rate". With such an incomplete control it would be unable to induce the first best optimal growth path, but would aim at a second best: in this respect the problem would be similar to the one discussed in Part 5 of our Chapter 5 where the government was assumed to control only public investment. We would now have to consider whatever particular path of the tax rate would minimize the loss with respect to the first best, the loss being measured on the intertemporal utility of the representative individual. This loss with respect to the first best would result from the distortions unavoidably induced by the tax system. Intuition suggests that, if exogenous elements of the general equilibrium model, other than public expenditures, are invariant over time, the second best path of the tax rate so found will be smooth because any other feasible path for financing the same flow of public expenditures would have to lead to a lower utility, hence to induce "additional tax distortions".

Rather than building such a second best argument in a general equilibrium model[283], Barro assumed that the social planner was able to directly find the second best path of the tax rate τ_t because the loss in utility with respect to the

[282] R. Barro, On the determination of the public debt, *Journal of Political Economy* (October 1979).

[283] The long footnote 7 of Barro's article describes such a model in which his simple result applies. The last sentence speculates about more general models in which the result would approximately apply.

first best could be given the following particularly simple expression:

$$\sum_{t=1}^{\infty} \beta_t \tau_t y_t \varphi(\tau_t), \tag{340}$$

β_t being the same real discount factor as is appearing in the intertemporal budget constraint, which may be written in our notation as:

$$\sum_{t=1}^{\infty} \beta_t (\tau_t - x_t) y_t = \frac{B_1}{p_1} \tag{341}$$

(with respect to (337) and (339) seigniorage is ignored). The function $\tau_t \varphi(\tau_t)$ of τ_t is assumed increasing and convex: the real loss due to tax distortions is assumed proportional to output as well as increasing at an increasing speed with respect to the tax rate. It is then immediate to conclude that, if government controls only the tax rates and has to finance the exogenous flow of expenditures $x_t y_t$, then in order to minimize (340) subject to (341), it must choose the constant tax rate $\tau_t = \tau$, where the value of τ is imposed by the constraint (341).

Accepting the simple assumption made by Barro we conclude that contemplated future tax rates have to be all equal: this amounts to full smoothing. If an exogenous shock occurs at $t = 1$, leading to an increase in x_t the present and future tax rates τ_t will have to increase; but they will change little if the shock is perceived to be temporary. On the other hand, if it is perceived to be permanent, tax rates will have to increase by the exact amount of the increase in x_1.

(iii) An empirical literature studies whether tax smoothing, as recommended by the above normative theory of public finances, appears as a positive feature in historical series of tax rates, public expenditures and public debt. The tax-smoothing property must then be tested within a formal representation of the constraints and exogenous elements imposed on the process linking these variables. Rigorous and appropriate tests are more difficult to design than we might have thought because, for instance, the time-profile of expenditures, if it is really taken as exogenous, cannot be closely approximated by a simple stochastic process: wars mean major breaks from previous movement; over the last century the share of expenditures in GDP during peaceful times systematically increased, this being particularly fast in some periods.

But, without closely looking at econometric tests, we may accept the following qualitative conclusion given by A. Alesina and G. Tabellini[284]: historical

[284] A. Alesina and G. Tabellini, Positive and normative theories of public debt and inflation in historical perspective, *European Economic Review* (April 1992).

time series are "approximately consistent [with the tax-smoothing theory], particularly in the UK and US. Two general features of the data account for this conclusion. First, the largest public debt issues occur during wars, which are temporary. Second, most shocks to government spending or transfers during peaceful times are mainly permanent, and tax rates tend to vary almost one for one with expenditures over time".

The fit is, however, not close. There are cases in which credible explanations of some large increases in the debt by tax smoothing do not seem to exist: such is the case with the increase in the US public debt during the 1980s. Still more clearly, as we noted earlier, the records show that often the debt-to-GDP ratio increased in certain countries but not in others, whereas all were otherwise experiencing about the same conditions. It is then hard to dismiss political reasons from the explanation.

4.7.4. *Some political economy arguments*

In Section 3.8 of Chapter 8, ending our discussion about the credibility of policies and the reputation of policy makers in the fight against inflation, we put a limit to the domain covered in this book at the frontier between macroeconomics and political economy. We announced that we would be at the frontier again "at the end of Chapter 9". We are indeed now led there by our discussion of public debt management. We are even going to briefly enter political economy, relying for the purpose on a survey by A. Alesina and R. Perotti[285], which precisely deals with the subject of this section.

The survey is organized around two questions: "why did certain OECD countries, but not others, accumulate large public debts? Why did these fiscal imbalances appear in the last twenty years [1975–94] rather than sooner?" The authors first argue that the "public choice" school does not provide convincing answers to these questions because it has to rely on ad hoc statements about a systematic bias in the fiscal illusion of voters whom governments were attempting to please (and to confuse) by raising spending more than taxes. Alesina and Perotti then turn their attention to the fact that public debt management is an important component of the political strategy of the government in office.

The stock of debt links past policies to future policies. The current government can affect the state of the world inherited by its successors; this strategic advantage can lead to an accumulation of government debt beyond the optimal level prescribed by tax-smoothing models. For instance, with uncertainty about

[285] A. Alesina and R. Perotti, The political economy of budget deficits, *IMF Staff Papers* (March 1995).

the preferences of future majorities over the composition of spending (as between "defense" and "social welfare", say), each government chooses to issue debt in order to limit the freedom of its successors and so to tilt the future composition of spending in its favour. The phenomenon will be strongest, hence the public debt highest, where and when political instability will also be highest. A cursory examination of empirical evidence suggests that the correlation is indeed real.

Political arguments can also explain why unavoidable tax rises or expenditures cuts are often delayed. If a manifestly permanent shock just perturbed the government budget so that a deficit appears and the debt begins to accumulate, a social planner would react immediately in order to maintain the balance of public finance. But in fact a distributional struggle begins among social groups and the parties aiming at representing them. The stakes are then to know what kind of expenditure ought to be cut, or for which tax should the rate be raised, or perhaps whether an increase in inflation might be accepted in order to provide extra seigniorage. The distributional effects differ. Those who are going to be the main losers object in each case to the move that would hurt them. The political parties most concerned, which may be part of a coalition government in office, organize demonstrations or threaten to leave government. None accepts to give in when pressed by others. Such conflicts, in which each player strives at forcing others into accepting sacrifice, decisions being so delayed, are known as "wars of attrition" in the theory of games, which discusses what the interactive rational solution might be.

Research in political economy stressed that the higher the number of parties in a coalition government, the higher is public debt. It also appears that longer-lived governments have smaller deficits. In other words, fragmentation of power leads to myopic policies, such as borrowing, delaying a tax reform or tolerating inflation. Research went deeper, because the degree of political heterogeneity within government or durability of governments depend on such institutional features as the party structure and the electoral system that generated it.

Empirical evidence coming from industrial countries since World war II shows that "the high public debt countries are almost exclusively parliamentary democracies with a highly proportional electoral system, and conversely almost all countries with such a political constitution have very high public debt. As a consequence of their constitution, all the high debt countries have very unstable governments, generally formed by a coalition of parties. Finally, while there is no... evidence of [correlation over time between] debt and inflation within each country, all the countries that collect large amounts of seigniorage also have high and explosive public debts... Similar evidence comes from developing countries"[286].

Research also brought out that budgetary institutions, which cannot be changed as easily and frequently as the budget itself, have effects on fiscal policy outcomes. All the rules and regulations according to which budgets are drafted, approved and implemented matter in order to know whether control is strict and unified or loose and scattered. Such rules and regulations can actually explain cross-country variations in deficits and debts, because they greatly vary across countries. From an in-depth study of this variation among the twelve countries of the European Economic Community, J. von Hagen[287] derived strong support to the following structural hypothesis: "Budget procedures lead to greater fiscal discipline if they give strong prerogative to the prime minister or the finance minister; limit universalism, reciprocity and parliamentary amendments; and facilitate strict execution of the budget law". In this statement "universalism" is defined as the property of a budget that includes "something for everybody"; "reciprocity" is an agreement not to oppose another representative's proposal in exchange for the same favour.

Clearly these various contributions of political economy are instructive for macroeconomists. They confirm the idea that political incentives create a bias towards both high debt and high inflation. Anyone with some practice of macroeconomic policy advising was aware of constraints imposed by the political and institutional environment in which he or she had to work. But making such awareness less anecdotal and more systematic is contributing to the multifarious knowledge thanks to which macroeconomic policies can become more effective.

4.7.5. *Effectiveness of fiscal policies in the short run*

Up to now the discussion of this section focused on the long-run aspects of fiscal policies or on government responses to the need for fiscal consolidation. No attention was given to the postwar Keynesian view of fiscal policies aimed at short-term regulation of aggregate demand. This view had been an explicit motivation of our treatment of the short-term equilibrium in Chapter 7, not the only motivation however. Starting with Chapter 8 the study of monetary policies as the instrument for short-run regulation held a central place, so displacing the earlier interest in fiscal policies. The various references of this chapter

[286] A. Alesina and G. Tabellini, Positive and normative theories of public debt and inflation in historical perspective, *European Economic Review* (May 1992), p. 342.

[287] J. von Hagen, Budgeting procedures and fiscal performance in the European Community, University of Mannheim, and Center for German and European Studies, University of California, Berkeley (August 1992).

to such policies were rather methodological or incidental. We may therefore feel the need for a clarification about the stabilizing role that macroeconomists should now assign to fiscal policy.

Possible discussion does not bear on the passive role: as we saw, most macroeconomists accept the idea that budgetary authorities should not interfere with the normal play of automatic stabilizers. But orthodox Keynesianism was asking for more: budgetary authorities had to make a frequent diagnosis about the trend of aggregate demand and to engineer corrective fiscal pushes or restraints. Did confidence in such an active role disappear?

It is probably so in the US, on the ground that the Treasury cannot be an active player in practice, given the budgetary institutions of the country. Declarations of US macroeconomists look almost unanimous in this respect. The reader may find not only a proof of this statement but also matter for reflection in interviews published by B. Snowdon and H. Vane[288]. For instance, to the question "Does it mean that the discretionary stabilizing role of fiscal policy... is dead?", R. Solow answered: "I am going to state an American's view about America; but I think at least some of it applies to Europe as well. Discretionary fiscal policy seems to be paralysed for two reasons. First,... fiscal policy has become a one-way street [to indiscriminate tax reduction]... But intelligent fiscal stabilization has to be able to move both ways, roughly symmetrically. Secondly, it is hard to invent and [long to] negotiate a neutral policy package...: timely fiscal stabilization seems to be very unlikely... So monetary policy gets the stabilization job by default" (p. 286).

A quotation of M. Friedman is also interesting, considering his role as a critic of Keynesianism throughout decades. To the question "What role do you see for fiscal policy in a macroeconomic context?" he answered: "None... I'm expressing a minority view here but it's my belief that fiscal policy is not an effective instrument for controlling short-term movements in the economy... I have never found a case in which fiscal policy dominated monetary policy... There are two possible explanations for that. One... is that the Keynesian view that a government deficit is stimulating is simply wrong... But there is the other reason that it is much harder to adjust fiscal policy in a sensitive short-term way than it is to adjust monetary policy" (p. 138). (Note that, in this quotation, Friedman accepts the idea of an active monetary policy for short-run stabilization).

However, depending as it is in a number of interviews on the US budgetary institutions, the opinion commonly expressed by American economists has

[288] B. Snowdon and H. Vane, *Conversations with Leading Economists – Interpreting Modern Macroeconomics* (Edward Elgar, Cheltenham, 1999).

ambiguous implications for whoever works in a country endowed with other institutions. What can we say as applying to countries in which some macroeconomically significant budgetary decisions can in practice be taken and implemented within a rather short time span?

There is, of course, the difficulty of a timely and accurate diagnosis, a difficulty underestimated in the 1960s by those claiming to be experts in "fine-tuning". But it bears on monetary policy as well as on fiscal policy. It simply means that "it is not easy for policymakers to find an appropriate" window of opportunity "for discretionary policy actions"[289]. But once such a window has been identified (as it could at times be, for instance, in Europe in late 1992), why could not fiscal stimulation or restraint be used, probably in conjunction with a monetary action? Indeed discretionary fiscal moves should always be considered, rather than dismissed off-hand, in cases of that sort.

However, developments in macroeconomic theory and, much more, experience of macroeconomic policies lead us now to immediately ask: how will public confidence be affected? An induced change in public confidence will also change the macroeconomic impact, sometimes for the better, sometimes for the worse. Moreover, reactions of public confidence, although sometimes surprising, are not wholly unpredictable. The art of analysts working in policy advising is to make good predictions in this respect.

But art is assisted by knowledge, even by scientific knowledge. In the matter there seems to be enough systematic traits for such a knowledge to build up. Two factors seem to be relevant for predicting how public confidence will react to discretionary fiscal action: the importance of the public debt, the credibility of government's announcements of its macroeconomic policy.

When the debt-to-GDP ratio is perceived to be too high, *a fortiori* when it moreover increases fast from one year to the next, the chance is high that any discretionary fiscal stimulus meant to counteract a depressive trend be doomed to failure: public confidence will deteriorate, which will lead to a decrease in investment demand and increases in savings. The risk that such a situation be faced in the future is precisely the reason why active fiscal stabilization must be patterned in such a way as to "move both ways", and the reason why the concern for public confidence imposes a stricter bound to the debt-to-GDP ratio than consideration of long-run solvency alone would imply. We also note that the effect here in question explains why strong fiscal contractions may in some circumstances turn out to be stimulating aggregate demand (see Section 7.1 of Chapter 7).

289 Page 626 in A. Lindbeck, The West European employment problem, *Weltwirtschaftliches Archiv* **132** (1996), Heft 4.

A credible announcement by government of an active temporary fiscal move may also contribute to improvement in public confidence. Indeed, consider the case in which the diagnosis of a depressing spontaneous trend is accepted. Public confidence deteriorates, particularly if government does not act. But, where public debt is already too high, announcement of a substantial fiscal stimulus will not only increase worries among those who are concerned by the long-run implications, but also more generally appear hardly credible, the suspicion being that no substantial stimulus will in fact be implemented: public opinion will interpret the announcement as a recognition by government that it is unable to do anything. In contrast, where public debt is moderate, a similar announcement is likely to be understood as meaning that authorities are coping with a difficult situation. If so, public confidence will be somewhat restored and private aggregate demand be sustained before any effect on private incomes is felt.

4.7.6. *The ethics of macroeconomic policy advising*

Acting as an adviser may at times set economists delicate ethical problems. This may occur, although in different ways, whether the economists in question work for the government or for an institution the main function of which is to inform public opinion. Let us begin with the first case.

As we saw in the conclusion of Subsection 4.6.1, economic advisers act as intermediaries between those supposed to know and those supposed to act. Anyone of these intermediaries has the special function to serve a policy maker by giving him or her objective assessments of the effects of alternative policies. Advisers cannot ignore that they are so taking part in a political process, but with a politically minor role. Since the role is so minor, many advisers felt no moral embarrassment to keep working in the same function after a change in government following a change in the political majority of the country: wherever advising a government is not perceived to be giving it a political support, there does not seem to be an ethical problem. But there are of course limits to what an adviser may so accept when he or she estimates that some decisions are morally wrong. In such cases the ethical problem may, however, appear less serious and distressing than the one faced by civil servants who have to implement the decisions.

Given their function as intermediaries advisers have secrecy obligations: they must not disclose information that the government is keeping secret, in particular they must not reveal what are contemplated economic policies on which they are working. Contrary to what is sometimes hinted, such a principle is not objectionable (again, subject to the proviso stipulated at the end of the preceding paragraph): nobody expects full transparency of the preparatory reflections pursued at the ministerial level; it belongs to ministers and not to their advisers to decide what should be disclosed. As we shall see in a moment, the ethics of whether to disclose or not may in contrast be difficult for those working in institutes whose main function is to enlighten public opinion.

In practice the main ethical problems faced by government advisers rather comes from the limitation or uncertainty of their knowledge (and this applies as well in other fields than economics). Strictly speaking, objective assessments would require perfect knowledge. But then in macroeconomics so few assessments would meet this strict requirement that advisers would be paralysed if they were not softening it thanks to a good deal of tolerance and judgement.

To appreciate the real dilemma one has to refer to the respective dominant psychologies of policy makers and researchers. The typical profile of politicians implies that they are eager to act and spend energy in finding the political arrangements which will lead to the aims they want to reach. Politicians are typically impatient about non-political "technicalities". In some cases they are not even listening, when the adviser tries to explain the complexities of the situation, the uncertainty of the diagnosis or the fact that various effects of a contemplated decision will go in opposite directions and that their respective strength cannot be precisely assessed. Hardly any decision maker was trained in strategic decision making under uncertainty, although some have intuitive understanding of the stakes.

All this means that the adviser has to be very selective in his or her choice of the difficulties to be reported to the decision maker; for the rest he or she has to directly decide on how to simplify and use certainty equivalents, or to judge which effects will dominate and which can be ignored. If he or she fails to make such a intensive screening of the issues, the likely outcome is that the decision taker will behave as if the economic adviser had nothing to say. The economic aspects of the problem will then be overlooked or, often worse, seen through ideological prejudices, fashionable popular economic writings and the like. Clearly most academics are not pleased by this requirement to decide what to select and then to speak as advocates in order to explain their recommendations: such a behaviour means to hide doubts, which the psychology of research in contrast stresses.

Ethical dilemmas then arise in a context where, on the one hand, preliminary decisions and uses of judgement on the part of the adviser forced to simplify are not congenial to the spirit of research, on the other hand, the early screening of issues may be determinant when the political decision will be taken. The risk exists that the adviser goes beyond his or her role, presenting as objective knowledge something which has little scientific backing and reflects more the fragile fashion of the day, or something which is already loaded with hidden value judgements, perhaps at variance with the values openly praised by the government. How far should the adviser go in simplifying? Far enough for being understood, but not too far, beyond what the function assumes. In practice knowing what this principle implies in each case is not obvious: here is the main source of the ethical problems confronting advisers working for political authorities.

This source of dilemmas also exists for analysts of macroeconomic policies working in institutions devoted to enlighten public opinion. But the dilemmas are less acute. Proper simplifications have also to be found. But the information needs not go up to precise policy recommendations. Enlightening analyses are usually meant to serve a less immediate purpose: to contribute to the evolution of ideas and public knowledge. The domain of each analysis may be clearly defined. Each publication complements earlier ones and will be complemented by others.

On the other hand, we may wonder whether such macroeconomic analysts are not facing ethical problems concerning the appropriate degree and form of self-restraint in the diffusion of information about a mounting problem which may turn out to be more or less difficult depending on how public opinion will react to it: for instance, how should analysts write about a recently identified and developing depression? Would not public confidence be strongly damaged if the analysts insisted heavily on the forthcoming difficulties? Since we recognize the importance that public confidence may at times have for the course of the macroeconomy, we can hardly discard responsibility of all others than government.

Author Index (Volume 35A&B&C)

This Page Intentionally Left Blank

Subject Index (Volume 35A&B&C)