contributions to economic analysis

János KORNAI

Economics of Shortage

Volume A

North-Holland

ECONOMICS OF SHORTAGE
VOLUME A

CONTRIBUTIONS TO ECONOMIC ANALYSIS

131

NORTH-HOLLAND PUBLISHING COMPANY

AMSTERDAM · NEW YORK · OXFORD

ECONOMICS OF SHORTAGE
Volume A

JÁNOS KORNAI
Professor of Economics
Institute of Economics
Hungarian Academy of Sciences

1980

NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM · NEW YORK · OXFORD

ISBN for Volume A: 0 444 85426 6
ISBN for Volume B: 0 444 86058 4
ISBN for both volumes: 0 444 86059 2

Publishers:
NORTH-HOLLAND PUBLISHING COMPANY
AMSTERDAM • NEW YORK • OXFORD

Sole distributors for the U.S.A. and Canada:

ELSEVIER NORTH-HOLLAND INC.
52 VANDERBILT AVENUE
NEW YORK, N.Y. 10017

Library of Congress Cataloging in Publication Data

Kornai, János.
Economics of shortage.

(Contributions to economic analysis ; 131)
Includes bibliographical references and index.
1. Scarcity. 2. Europe, Eastern--Economic conditions. 3. Hungary--Economic conditions--1968- I. Title. II. Series.
HC244.K66713 1980 330.947 80-19089
ISBN 0-444-85426-6

Printed in the Netherlands

Introduction to the series

This series consists of a number of hitherto unpublished studies, which are introduced by the editors in the belief that they represent fresh contributions to economic science.

The term 'economic analysis' as used in the title of the series has been adopted because it covers both the activities of the theoretical economist and the research worker.

Although the analytical methods used by the various contributors are not the same, they are nevertheless conditioned by the common origin of their studies, namely theoretical problems encountered in practical research. Since for this reason, business cycle research and national accounting, research work on behalf of economic policy, and problems of planning are the main sources of the subjects dealt with, they necessarily determine the manner of approach adopted by the authors. Their methods tend to be 'practical' in the sense of not being too far remote from application to actual economic conditions. In addition they are quantitative rather than qualitative.

It is the hope of the editors that the publication of these studies will help to stimulate the exchange of scientific information and to reinforce international cooperation in the field of economics.

The Editors

Contents

Volume B

Foreword

In 1976–77 I spent fifteen months in Sweden at the invitation of the Institute for International Economic Studies of Stockholm University. My work included a series of lectures held under the title "Economics of Shortage". The present book has grown out of the notes of these lectures.

I have learned a lot from the discussions and arguments I had with the audience at my lectures, with my colleagues working at the Institute, and members of other universities and research institutes. I wish to express my thanks for stimulating questions and valuable remarks particularly to Åke E. Anderson, Rolf Eidem, Sven Grassman, Pentti Kouri, Harvey Lapan, Clark Leith, Assar Lindbeck, Erik Lundberg, Lars E. O. Svensson, Jörgen W. Weibull, and Bengt-Christian Ysander.

I remember with gratitude my Swedish hosts. The problems treated in my book became known to me while living in Hungary and participating actively in events. But a temporary leave and the distance away from my country made it easier for me to consider those problems thoroughly. The generous support of the Stockholm Institute allowed me an extended period to concentrate my efforts exclusively on writing this book.

I avail myself of this opportunity to express my thanks to all who helped me at home in Budapest in the final completion of the manuscript with editorial and technical assistance; in particular to Mariann Dicker, Zsuzsa Kapitány, Maria Lackó, Edit Makó, Péter Mihályi, Judit Szabó, and Teca Zimányi.

Ilona Lukács and György Hajdu cooperated with me in translating the book into the English language. The bulk of the work fell on Ilona Lukács. I am grateful for their care and devotion. I am indebted to Professor Paul Hare and Martin Cave (Stirling University, United Kingdom) for their extremely valuable help in correcting the English of the book.

Stockholm–Budapest, 1977–78 János Kornai

Notation

Symbols of real numbers are in italics, symbols of vectors in bold face italic type. An asterisk (*) added to the symbol of a variable denotes the *normal value* of the variable. A "hat" ˆ over a symbol denotes an *upper bound.*

If the same concept occurs in several chapters of the book, it is given the same notation throughout. For these concepts parts of the Latin and Greek alphabets are used; their list is given below. The other Latin and Greek letters denote different concepts in different chapters of the book, but only such concepts as occur exclusively in an individual chapter. Even in these cases a particular symbol denotes the same concept throughout any one chapter.

Some of the symbols in the list denote concretely defined concepts. Others serve to denote *families of concepts*. For example, all kinds of shortage indicator are denoted by z. In individual chapters some category of shortage indicators is also defined in concrete terms, but the letter z will always be used as the symbol for any kind of shortage indicator. (Perhaps as a basic symbol, complemented by special notation, e.g. z_L is the symbol for indicators of labor shortage.)

The expression $\boldsymbol{x} \geqq \boldsymbol{0}$ means that all components of vector $\boldsymbol{x}$ are non-negative. The meaning of the expression $\boldsymbol{x} \geq \boldsymbol{0}$ is that all components of vector $\boldsymbol{x}$ are non-negative, but at least one component is positive.

We shall use the following abbreviated notation:

$$(x)_+ = \begin{cases} x, & \text{if } x \geqq 0, \\ 0, & \text{if } x < 0, \end{cases}$$

a = input coefficient,
c = cost,
d = demand,
g = consumption,
p = price,
P = price index,
q = slack,

s = supply,
t = time,
u = output stock,
v = input stock,
w = friction,
x = production,
y = trade, turnover,
z = shortage,
β = degree of the softness–hardness of the budget constraint,
ζ = rigidity parameter,
κ = degree of capacity utilization,
μ = forced substitution propensity,
ρ = resistance parameter,
τ = delay time.

CHAPTER 1

Introduction

1.1. Outline of the main subject

The word "shortage" in the title of the present book is a summary denomination of a large group of phenomena . What comes to one's mind on hearing this word is shortage in consumer goods or housing; or queuing up in the strict sense of the word in front of the baker's or the butcher's shop; it may also be notional queuing, such as waiting for the installation of a telephone or admission to a hospital. Similar phenomena in production are labor shortage, shortage of material, of parts, or of electric current.

A series of economic phenomena are obviously related to the ones enumerated above: tensions on the market for investment goods since their demand usually exceeds supply; a chronic deficit in the balance of payments and the concomitant shortage of foreign currency.

There are numerous denominations in use in economic science to explain the whole group of phenomena or at least one or other of their significant features. The most important ones are the following: "shortage economy" (translation of the German "Mangelwirtschaft"); "seller's market"; "repressed inflation"; "taut planning"; "overambitious planning"; and "overheating" of the economy. The author of the present book has introduced two further expressions: "rush" in growth and "suction" on the market. These denominations are not all fully identical; they are not simply synonyms. Yet their listing together will obviously give rise to associations on the part of the reader.

In this chapter I do not strive to define exactly the concept of "shortage". I hope that first the examples, then the references to the many kinds of denominations used in the literature will be sufficient to give an idea of the subject of the book

It may not be necessary to dwell long on the *importance* of shortage phenomena. They can affect the everyday life of people in a hundred different ways. Shortages can upset the consumer's purchasing plans and frequently cause difficulties in consumer goods' supply. Much leisure time is lost waiting in the doctor's surgery or in a shop. It is no exaggeration to say that housing shortage can vitally affect a family, or that restrictions on admission to higher education can seriously affect the future of a young man at the beginning of his career. And, of course, one experiences the consequences of shortage not only as consumer but also as producer. The director, the foreman, and the worker repeatedly find themselves faced with troubles caused by shortages of materials or of labor, with all the accompanying stress. Shortages, in the final analysis, affect human relationships: the relations of consumer and producer, seller and buyer, the person deciding about rationing, and those receiving the rations.

Shortages are, either as the cause of other phenomena or as their consequence, connected by a million ties to other components of the economic system: to prices and to wages, to planning and to market, to fiscal and monetary policy, to material and moral incentives.

Although the importance of shortage has long been recognized by economists, and it is treated in thousands of books and studies, I have no knowledge of a book that would have this exclusively for its subject. The present work is the first attempt at comprehensively reviewing the economic problems of shortage.

Some of the shortage phenomena outlined above may appear in any kind of economic system. A number of them are not unknown in advanced capitalist countries, particularly at times of booms or in wars. At the present time the energy shortage is becoming more intense in a number of capitalist countries. Similar difficulties are also expected for the future. Shortage phenomena appear in various forms in nonsocialist developing countries. In this book, however, I do not deal with these countries: my study is centered on *socialist* economic systems.

I wish to stress that my work does not aim at elaborating the *general* theory of socialist economy; I discuss the versions of socialist economy that have been established in Eastern Europe. I consider mainly the small Eastern European countries. Although these are similar to the Soviet economy, the latter also displays special features which are explained by the huge dimensions of the Soviet Union. My book does not deal with these special features. Yugoslavia is also omitted from the examination because its system is essentially different from that of other Eastern European countries. I am not sufficiently acquainted with the socialist countries of Asia, Africa, and Latin America. I cannot take a stand on the question of whether similar phenomena occur there as in Eastern European countries.

It will be obvious to a specialist well acquainted with Eastern European countries that the book has been written by a *Hungarian* author, inspired by experiences gathered in Hungary. Although, of course, I do not deny this, I do strive to do more than research exclusively for Hungary. The "socialist economy" that is treated in this book is a "stylized" model abstracted from the existing Hungarian or Polish or Bulgarian particularities.

If, on the one hand, I stress the abstract nature of the model, I must add, on the other, that it is intended to describe a few features of historically developed and *existing* Eastern European systems in a general form. I do not deal with hypothetical socialist systems, or with the question of what the socialist economy would be like if it functioned otherwise than as it does.

If, in what follows, the book uses the term "socialist economy" without any distinguishing attribute, it must always be interpreted in the abovementioned *restricted* sense: it is always the abstract model of socialist economies actually functioning in Eastern Europe (except Yugoslavia).

Even though we have thus narrowed down the sphere of examination, further distinctions will still be necessary. In Eastern Europe profound changes have been – and are still – taking place in the institutions and in the management systems of socialist economy. This tendency shows itself with particular clarity in Hungary: since the 1968 reform, autonomy of firms has grown, plan directives for current production have been eliminated, and the role of markets has increased. In some of the later analysis it will be specified whether they are concerned with the *traditional* socialist economic management system prior to the reform, or with the *postreform* system. The book is concerned mostly with the former and does not undertake a detailed description of the situation that established itself after the reform. If, however, the situation after the reform is touched upon, it is always done with regard to the Hungarian experience up to 1978.

It may turn out that some of my propositions are also valid in a wider sphere and that they can be applied to nonsocialist systems as well. In a few places I comment, even though briefly, on one or other feature of the capitalist economy, mainly for the sake of comparison. The methodology, the formalism, and some of the suggestions on observation and measurement may also prove to be useful for the examination of such systems as are not characterized by shortage but its opposite: unemployment and underutilization of resources.[1] When I come to such results, pointing beyond the study of a socialist economy, I shall regard them as a bonus. The immediate objective of my efforts is, however, more modest than that. I would like to throw some light upon shortage phenomena in a socialist economy (first of all in the Hungarian economy), their causes as well as consequences.

Whoever glances through the contents of the book may think he is holding in his hand a comprehensive manual on socialist economy. That is because we shall cover almost every important chapter of economics: micro- and macroeconomics, short- and long-term decisions; demand, supply, prices, wages, employment and the role of money; households and firms. That is true – but all *only from the aspect of shortage*. For example, talking about prices the book does not give a general price theory, nor a comprehensive empirical description of the actual dynamics of prices, but tells only that much about prices as is closely related to the shortage problem. Other questions will be treated in the same way.

I can illustrate the foregoing by the following example. Let us imagine a book of one thousand pages that treats shortage as it appears in all kinds of economic systems. And let us imagine another book, say, of two thousand pages that explains every basic question of a socialist economy. The present study then contains the six hundred pages that are common to the books of one thousand and two thousand pages. That is to say, my subject is the overlapping or common part of two large sets of problems ("general economics of shortage" and "economics of socialism"). I ask the reader not to demand – whether he agrees or not with these six hundred pages – the two thousand four hundred pages that this book has no intention of discussing.

One basic problem of the capitalist economy is to achieve full employment. In the great 1929 crisis this question became acute in the extreme. That is why economists studying capitalism inevitably *had* to turn their attention to unemployment and the underutilization of resources. However

[1]Different chapters, and within any one chapter even different sections, may differ from each other as regards degrees of generalization. The book carries a few entirely general analyses covering *every* economy. In other places we have in mind the abstract model of the *Eastern European socialist* economy without historical and national distinction. In other places again we discuss exclusively the *traditional* state as before the reform, or only the *postreform* state, the latter perhaps based only on Hungarian experience.

This ascent and descent into higher and lower levels of generalization is inevitable. We try to make it easier for the reader by drawing attention in each case to the sphere covered by the analysis.

great the contributions of a few outstanding scientists, such as Keynes, and before him the Swedish School and Kalecki, it was the pressure of social problems that brought the subject to the center of attention. Although in the years following the Second World War it seemed to many that capitalism had finally overcome this difficulty, it has become topical again, and thus once more a central subject of economics examining capitalism. The problem of a socialist economy – if in this chapter I may put the point in a somewhat oversimplified way – is just of the reverse sign. That is, instead of underutilization, "overheating"; instead of the comparatively low level of aggregate demand, its comparatively high level; instead of unemployment, labor shortage; and so on. Shortage plays a similarly central role in understanding a number of problems of a socialist economy, as does the analysis of unemployment in studying capitalism. It is the present real problems of the socialist society that compel the researchers on the socialist economy to elaborate the theory in order to explain this group of phenomena.

1.2. Description, methodology, explanation

I consider it my first task to describe the *phenomena* themselves. Those who experience the symptoms of shortage directly and every day may oversimplify one or other aspect of them. And those who live in another social system and approach the subject from the outside, e.g. relying on superficial "tourist's experience", or on reading a short chapter about socialism in a comprehensive Western manual, usually have biased stereotypes in their mind. It is my aim to offer a more varied and complete *description* to replace those poor and simple patterns.

For this an adequate apparatus is needed. The primary subject of the book is *shortage*, and its related secondary subject – almost of equal importance – is the *methodology of economic systems theory and economic control theory*. Which are the models – mathematically formalized, visual or verbal models – by which to describe the control mechanisms, signal systems, and behavioral regularities of the various systems? Which are the variables and constants characteristic of the system; what observation and measurement possibilities are at hand? A great many things belong to the sphere of "methodology", and it would be easier to discuss them with the reader *after* he has read the book, than it is now when he has not yet become acquainted with the conceptual apparatus.

I would like to insert here a personal remark. It caused me considerable mental discomfort to find that some of the phenomena that faced me during the analysis of shortage could not be described adequately by the usual tools. The research worker tries first to squeeze reality into traditional, ready-made forms, and then something always hangs out. And, sooner or later, it will be more useful to adapt the apparatus to reality than reality to the apparatus. This fight between reality and the describing apparatus accounts for the fact that there is, inevitably, much methodological discussion in the book.

Description of the phenomena, and elaboration of the methodology to be used for description and analysis is, of course, only the first step. A deeper layer of examination is the elaboration of an *explanatory theory*. How are the separately described shortage phenomena interrelated? Is it by chance that they appear together, or are they inseparable? Are they occasional events, or are there regularities behind the random fluctuations? Within theoretical clarification it has to be clearly stated: what is the result of current *economic policy*; what comes from the *socialist economic system* or its various concrete versions and, finally, what are the phenomena to be found *in every system*? The causal analysis will evolve gradually, step by step before the reader.

To this, however, I must add: revealing the *causes* of shortage is not the exclusive purpose of this book. It is no less important for us to seek an answer to the question of how shortage *affects* the system. How is the economy functioning in the presence of chronic shortage?

Some of the ideas set forth in my book have already been discussed in the literature, or "by word of mouth": in the conversations of economists and managers. My primary aim was the *integration* of fragments of thought relevant to shortage. And, while striving for such a synthesis, I was obliged to compromise on a fine elaboration of details. Let me quote, at least so as to reassure myself, the words of the great Hungarian writer and thinker László Németh:

> There are two ways to think about science: it may be a building that has to be compiled brick by brick, and in which every little brick is valuable if it fits; or it may be a great explanation of life and the world, an answer to some anguish renewed in every generation, to which a full "reassurance" based on the momentary store of knowledge is sought... No bricks have ever built a house; before the bricks should start on their way toward the house, the house must be virtually there; partial works can build but a heap of stones, if they are not preceded by the design, the intended house, the house itself.

I felt that an attempt must be made at laying down the design of the "house": a theory explaining shortage, even if the "bricks" (enough empirical observation and their thorough mathematical–statistical analysis; uniform formal apparatus, etc.) are not all made yet. It is worth suffering the loss of elaboration and maturity of details in exchange for the gain that may be brought by the *comprehensive* character of the train of thought.

The remainder of this chapter will point out one by one the missing "bricks"; the most characteristic limits and weaknesses of the book, in the author's own judgement. This is not because I hope to disarm my future critics by this "list of shortcomings", but rather because the reader should be properly informed.

1.3. Theory and formalization

I wish to elaborate the *theory* of shortage phenomena present in socialist economy. Yet does it deserve to be called a "theory"?

Two definitions of "theory" are spread among economists. According to one, theory is the general description of interdependencies and regularities of a group of phenomena and explanations of their causes. The other definition reserves the term "theory" for rigorously formalized and mathematically proved statements. Personally I accept the first definition, in which case this book claims to provide a theory. It is obviously open to discussion whether it stands as a theory. It is open to verification or rebuttal in both a logical and an empirical way. The effort itself is, however, undeniable: it is the intention of this book to describe the interdependencies of the sphere of shortage phenomena, to analyze their regularities, and to find their causes.

There are economists who feel it a virtue of their work if their train of thought is not formalized. I do not consider it a virtue, but as a weakness. It is true that some of the phenomena discussed in this book have already been described with the aid of formalized models. I may refer to a few works of my own, or to others carried out together with co-authors, or to studies by my colleagues in Budapest and in Stockholm.[2] And also I can refer to products of similar intellectual trends. Yet every model that can be mentioned on the subject is one-sided and partial. There is no uniform and

[2]The present book and the volume of studies by Kornai–Martos (1979a, b) are "joint products" of the same train of research. The latter deals with the theory of "quantity" adjustment taking place without price signals, with the aid of mathematical models. The reader can gain in fact a full picture of our work if he gets acquainted with both works.

synthetic formalized model of the *whole* of the theory. The various partial models, although with no contradiction or incompatibility among them, have been constructed using different kinds of formalism. Their formal synthesis has not yet been achieved.

This will certainly discourage some readers who may feel that it would have been better to go on with research until it takes its final and closed mathematical form. As far as I am concerned I do not think it is obligatory to wait for that higher degree of maturity. I am confident that thoughts, if at all correct, are apt to promote, in their present loose form, a more strict formulation at a later date. In the history of economic thought quite a number of examples can illustrate this order of succession.

And, such being the actual state of formalization of the theory, I have tried in this book to present less rather than more. The main part of the book is basically of a verbal character. In a few places figures and schemes are provided to facilitate the explanation. Elsewhere formulae are also presented, although more for expository purposes. They may serve as a compact expression of an interrelation or of a more intricate chain of ideas. Or they may help in explaining a definition or a measurement problem. Verbal explanations are also associated with these formulae. It has been my aim to make Chapters 2–22 (or at least most of them) accessible without difficulty even to the nonmathematical economist. There is more than one formula in the book which one could easily present in a more concise form for the mathematically trained reader. And yet we chose the more detailed form, since that enables us better to attach a verbal economic interpretation to each individual term of the formula.

Chapters 7 and 8 step over these limits when they explain a few ideas with the aid of mathematical models. I tried to give these too the simplest possible form, so as to enable the reader less versed in mathematical economics to follow them without difficulty. A detailed description of the models and mathematical proofs of the propositions are to be found either in the literature referred to in the chapters in question, or in the Mathematical Appendices at the end of this book, which were written for the mathematical economist reader.[3]

I refer at several points to Hungarian and foreign research works whose formalism can be used for an examination of the phenomena related to shortage. In this way the book can serve as an "annotated bibliography" for the student and research worker in mathematical economics.

[3]Co-authors of the Mathematical Appendices are Jörgen W. Weibull and András Simonovits.

1.4. Empirical support

I would like to make the book "redolent of life", and I wish that all who know from experience the phenomena described should feel my statements true. At the same time I must admit that I cannot unambiguously verify by means of available data most of the assertions in the book.

My book usually gets to the point of indicating such variables, parameters, and indicators of which it can be said – relying on logical argumentation – that they are observ*able* and measur*able*, yet most of which have not been recorded by economic statistics. And where there is some measuring, then the mathematical–statistical and econometric analysis has not yet been done very throughly.

I regard with envy those of my colleagues who can support their assertions by masses of data and verify their hypotheses by econometric tests. The dispute about inflation, for example, carried on among Western economists, has a vast statistical and econometric background. The data background to my own research topics and the level of mathematical–statistical analysis of the observations are incomparably lower.

What I have said in connection with formalization, I can also say now concerning data and econometric analysis: I felt that it would not be right to wait. The history of economic research and statistics shows that the order of succession is sometimes the following: first the economic conjecture is put down, and this is followed by observation, data collection, and mathematical–statistical analysis supporting or correcting the conjecture.

I had to be content with formulating only *empirically testable hypotheses* in most cases instead of empirically verified theories. It far exceeds one single researcher's energy to check every hypothesis. It is possible that others in my place would have contented themselves with setting up hypotheses on one or two questions, which they would then have tested with the utmost thoroughness. However, the aim I set for myself was both less and more than that. I have tried to analyze shortage in numerous kinds of interrelations, so that in the end I should arrive at a general theoretical analysis. In the course of this we arrive at a *large number* of propositions, even if in most cases only in a tentative, hypothetical formulation.

In any case, I have insisted in my own examination that I should only suggest hypotheses of which it can be said: they are empirically test*able*. And, wherever possible, I have also tried to indicate the ways in which the testing could be done.

1.5. Sources, literary background

The book comprises rather a wide sphere of problems, and each has a vast literature. I make no claim to know more than a small fraction of the literature, even though I have tried to absorb as much of it as possible. It may be useful to make a few preliminary remarks on the principles of references and sources used.

(1) I am unable to follow the otherwise approved and generally expected tradition, in which the discussion of a question is accompanied by a full survey of the whole history of the theory. Since my book touches on almost every chapter of economic theory, even if from the aspect of only one subject, a thorough review of the history of theory would in itself make up a book. It is already on the grounds of space that I could not undertake to do it. I try to make up somewhat for this shortcoming by indicating summary works, manuals, and comprehensive surveys in order to facilitate the reader's study.

(2) When it is clear to me exactly where some proposition comes from, I make reference to it. There are, however, a number of ideas for which I do not know who has priority; and I cannot undertake an investigation to clarify the origin of the idea.

(3) In most cases I refer to such works to support or complete what I have to say. With a few exceptions, I do not enter into polemics with other authors. I put the emphasis on the positive explication of my thoughts. To discuss how much this agrees with, or deviates from, the opinion of others, will be the task of later works.

(4) The literature of socialist countries is represented almost exclusively by Hungarian authors. It is not the language problem only that held me back, since some works written in other socialist countries are available in languages I understand. I have learned, however, from domestic experience that an article or book can be truly appreciated if one has an "inside knowledge" of the disputes among which it appears and, what is even more important, if one also has an "inside knowledge" about the economic reality that is reflected in those disputes. I would not like to pick one or other work "from the outside", that has by chance come into my hands, with almost inevitable arbitrariness. I hope that my book will reach the economists of other socialist countries, and then they will have to form an opinion as to whether my statements are also valid for their economy.

(5) My book has been written primarily for the theoretically well-versed reader. I presume some acquaintance with Marx and Keynes, and with the

dispute between Keynesians and anti-Keynesians, as well as some knowledge of neoclassical theory.

(6) I finished the correction of the manuscript of the book in the autumn of 1978; accordingly, I had studied the literature up to that date. The fact that there are in the list of references a few works published in 1979 is because I saw their manuscripts in 1978.

While my book relies largely on printed literature, I consider as a mental source of almost equal rank the professional "public opinion" of Hungarian economists and managers. This public opinion has affected me in different ways. Sometimes it was the actual *question* raised in the debates that inspired me; at other times it was the *answer* of one or other party that convinced me. On a few occasions I considered the answer to be final; on other occasions I considered it as a conjecture going in the right direction which I tried to formulate more clearly, i.e. to express it in theoretical language.

Therefore a number of statements will be found in the book to which my colleagues will respond: "This is not new; we have known this before." Although there may be some who will say so with some irony, I shall accept it gladly. This is exactly one of my aims: to express such thoughts as have occurred to many others – such thoughts as are "in the air". And to do it in such a way that they should compose a logically coherent system, i.e. construct a synthesis.

The writing of this book has been preceded by innumerable discussions through the years. I shall name here only a few of those whose information, analyses, or questioning greatly affected me. Among the functionaries of Hungarian economic life I shall mention Deputy Head of Department, János Árvay (Central Statistical Office), Head of Department, Tamás Bácskai (Hungarian National Bank), Deputy Head of Department, Ákos Balassa (National Planning Office), Deputy Director-General, Andrea Deák (State Development Bank), Deputy Chairman, József Drecin (National Planning Office), Minister, Lajos Faluvégi (Ministry of Finance), Secretary of State, István Hetényi (National Planning Office), and Director-General, László Szabó (National Market Research Institute). I feel that the many hours that the above-listed economic leaders and a number of others not mentioned here spent in discussions with me were of invaluable assistance to me in getting acquainted with Hungarian economic life.

Beside *officials and managers* of the economy I am also much indebted to *scientific researchers* for the discussions and debates I could conduct with them. In the first place I wish to mention those with whom for years I

have been carrying on research works in common at the Institute of Economics of the Hungarian Academy of Sciences: Katalin Farkas, Zsuzsa Kapitány, Mária Lackó, Béla Martos, András Simonovits, and Judit Szabó. Whenever I turned to them for advice or actual cooperation in research, I could always count on their helpfulness. I appreciate particularly the valuable spiritual support I received in the form of information, suggestions, and critical remarks from the following: Tamás Bauer, András Bródy, András Nagy, Tamás Nagy, Rezső Nyers, Judit Rimler, Attila K. Soós, Ágnes Ungvárszky (Institute of Economics of the Hungarian Academy of Sciences), Zsuzsa Dániel (Planning Institute, National Planning Office), János Gács, Mihály Laki, Márton Tardos (Market Research Institute), and Attila Chikán, Miklós Riesz, János Timár (Karl Marx University of Economics).

A number of those listed above – officials and managers as well as scientific research workers – read the first version of the manuscript. I received from them countless valuable remarks and advice which I tried to utilize in the final formulation of the text. May I avail myself of this opportunity to thank them for their help.

As already indicated by the length of the list of names this is really a wide sphere of professional "public opinion" that I try to express. But, of course, while emphasizing my own indebtedness, I must make it clear, too, that certainly many of those named in my list see things differently from me. The author alone must assume responsibility for the train of thought of the book, and for possible mistakes and errors.

With regard to literary sources, I would like to mention that I consider this book as a direct continuation of my earlier works *Overcentralization*, *Anti-Equilibrium*, and *Rush versus Harmonic Growth*.[4] In this connection I faced the following dilemma in writing the present book.

On the one hand I have to inform the reader about the relation of my present book to the earlier ones. On the other hand, as a reader I have an aversion to authors forever citing themselves, and I did not want to commit the same fault. Finally, I chose the following compromise. I shall refer in detail to my earlier works at such points where my present view is different from my earlier ones. (See Chapters 5, 7, 8, and 21.) In some places I shall make brief reference to one or other of my previous publications, if the empirical description in them completes and supports

[4]See Kornai (1957, 1959, 1971a, b, 1972a, b).
The present book has been preceded by a few partial studies. (See Kornai, 1974, 1975b, 1976a, and 1976b, as well as the works listed in footnote 10, Chapter 7.)

what I have to say here. On such occasions I wish to use the reference to avoid repeating myself. I do not, however, make any reference in cases where the present book uses and further develops theoretical and methodological ideas raised in earlier works, first of all in *Anti-Equilibrium*. The reader may be convinced – without proof of concrete references – that *Economics of Shortage* is in many ways related to *Anti-Equilibrium* and that it tries to fulfil a number of research tasks indicated in the earlier book.

1.6. A few more delimitations

I have so far delimited the subject of the book from several viewpoints; to such an extent that the question may even arise whether it is not an exaggeration to call it *Economics of Shortage*. I might count upon more indulgence in the matter of the title, too, if the Introduction reveals quite frankly the limits of this work. In what follows I briefly survey the subjects excluded from the book and not mentioned so far.

(1) The book has been formulated in such a way that it should also be possible for the foreign reader to follow. This is a logical consequence of the preliminaries mentioned in the Foreward: the book grew out of a series of lectures held before a foreign audience. Therefore such information must be given at certain points as may be unnecessary for the Hungarian reader. And further complementary reading will be needed by those who want to acquire a thorough knowledge of the *historical, social, and institutional background* of these problems. Fortunately, a rich literature is available, much of it published in foreign languages.[5]

(2) I shall touch on the *political* aspect of the problems in a number of fields. However, I do not consider it part of the subject of the book to analyze in detail the relationship between politics and the economy, or the role of the Party and of other political authorities in the control of the economy.

[5]From the literature treating the *Hungarian economic policy* of the recent past we shall name the following works: Berend (1974), Drecin–Hetényi (1970), Friss (1976a), Friss (1976b), Hetényi (1976), Jánossy (1969, 1970), Nyers (1978), and M. Timár (1975a, b).

Within literature dealing with the *Hungarian economic management reform* the following works are mentioned: Bálint (1970), Csikós-Nagy (1978a, b), Friss (1971), Gadó (1972, 1976a, b), Nyers (1969a, b), Péter (1956), K. Szabó (1964, 1975), and Tardos (1975a, b).

(3) The book deals comparatively little with *planning*. On that area also abundant literature is available.[6]

It is a well-known fact that an important role is played in a socialist economy by *vertical* relations between central, medium- and lower-level management authorities and firms as well as nonprofit institutions. The book deals with them in several places, e.g. Chapters 3, 5, 9, 13, and 22. In addition, however, it gives more scope to the examination of *horizontal* relations between firms, as well as between firms and the household, if only for the reason that the latter are less fully treated in the literature.

(4) We do not analyze the *organization* of production and trade. Thus, among other things, the discussion neglects such questions as which market structure (monopoly, oligopoly, imperfect competition) is prevailing in different markets; what is the degree of concentration of production and trade; and how all this affects shortage phenomena.

(5) The book approaches several questions from the *macro*economic viewpoint. (See, for example, Chapters 9, 11, 12, 16, 19, and 21.) Yet the larger part of the book is of *micro*economic character. Its main task is to clarify the micro-foundations of macro-processes.

(6) In a few places, e.g. in Chapters 9, 10, and 11, we analyze the *long-term* processes of the economy. Yet the book centers its attention mostly on *short-term* adjustment. It considers as given the resources, institutions, organizational forms, and control mechanisms of the system, and together with these the accustomed norms of behavior and characteristic reactions of decision-makers are given. The question is the following: how is the economy functioning in a given situation?

(7) I disregard *foreign trade* in general. Reference is made only sporadically, without any pretension to completeness, to foreign economic relations.

(8) I do not discuss the special problems of *agriculture*.

(9) When discussing firms we mean almost exclusively *state* firms. We do not discuss the particular problems of cooperatives. We shall only touch upon, in a few places of the book (e.g. in Chapters 11, 16, 17, and 19), the sphere that is called the "*second economy*" or "informal sector".[7] There

[6]From the most recent works see Augusztinovics (1979) and A. Balassa (1979). I was also induced to give a comparatively brief treatment of planning by the fact that in my books, Kornai (1957, 1959, 1973, 1975a), I dealt in detail with the subject. As I have just said, I have tried to repeat as little as possible from my earlier works.

On the international experience concerning mathematical planning, surveys are provided by Johansen (1977) and Blitzer–Clark–Taylor (1975).

[7]On this sphere of problems only a few Hungarian studies are available. Valuable analysis is to be found in the article by A. Hegedüs-Márkus (1974) and Gábor (1979), and in the paper by Gábor–Galasi (1978).

belongs here a wide variety of activities, from industrial and commercial private enterprises functioning with official permission through the "grey" and "black" market to various half-legal and illegal services.

One of the reasons for the restrictions under section 2.9 is the incompetence of the author. The other reason is simply that the subject of the book has many dimensions, and its volume – even though it has grown large in the end – is restricted. It is inevitable that we should omit a few subjects, however important and interesting they may be in themselves.

(10) At the end of the delimitation of the subject one more point has to be clarified. My book undertakes the elaboration of a *descriptive–explanatory theory*, and does not provide a *normative* theory.

From my analysis, if it proves to be correct, a number of practical economic policy conclusions can be drawn. The present book, however, does not exploit these; it does not contain practical suggestions as to what should be changed either in economic policy or in economic institutions. I made proposals in a number of my earlier works, and it is my intention to elaborate suggestions also in the future. In *this* book, however, I refrained from doing so even if the formulation of the suggestion was "on the tip of my tongue".

The deeper I tried to penetrate into the subject discussed in the book, the clearer it became to me what difficult and deeply rooted problems face us. It can easily turn out, with hastily drawn-up suggestions, that they can only bring about a half-solution, or may even cause the situation to deteriorate further. It would also be a constructive contribution to the development of economics if the "anatomy" of the existing situation is thoroughly worked out, that is, the deeper-lying reasons and interrelations of shortage phenomena are clarified. This could be the first step that may be followed by other, more practical developments.

PART I

ADJUSTMENT WITHOUT PRICES

Introduction to Part I

Three main types of micro-organizations in the economy will be discussed.

(1) The *firm* that sells its output for money and covers its expenditures exclusively or primarily from proceeds.

(2) The *nonprofit institution* which provides its goods or services free to the user. Most nonprofit institutions cover their expenses from money allotted by the state budget. We deal exclusively with this case. We disregard institutions that subsist on donations, and organizations maintained by membership fees.

(3) The *household* which covers its expenses of consumption from monetary receipts acquired through work or enjoyed by some other right.

Authors discussing shortage in the socialist economy usually start from the market for consumer goods or the experiences of the household sector. My book departs from this practice, its starting-point being the examination of the sector of firms. It is my conviction that the root of the problem is there: in the sphere of *production*.

Part I of the book deals almost exclusively with the sector of firms. In some places we refer briefly to phenomena appearing in the nonprofit institution sector. The household sector will be studied more thoroughly only in Part II. (An exception is Chapter 11 treating employment, which touches also on the problems of households in connection with labor supply.)

The firm's behavior is guided by many different motives; for example, it can be influenced by the motive of acquiring and increasing *profit*. The firm's decision is affected by numerous signals, such as *price*, of both input and output, and among input prices also the *wage*. The firm's actions may be constrained by various factors, for instance the amount of *money* available. In Part I we generally disregard *profit*, *price*, *wage and money*, except for some occasional references to their role. It is not only *changes* in prices that will be disregarded, but also – whether fixed or changing – all their effects on decision-makers. The detailed examination of the role of price, wage, profit, and money is left to Part II.

There are several considerations to account for the disregarding of prices and other related factors in Part I. The most important one is that in the traditional socialist economic management system their effect is in fact very weak, as will be shown in more detail in Part II. And while price, profit, and money have little effect on the firm, it is still obviously surviving, growing, and showing definite regularities in its functioning. Under such circumstances these regularities can be better understood if their theoretical analysis is carried out "purely", i.e. fully disregarding price and other related factors.[1]

This leads us to the second consideration accounting for the structure of the book. It has been clear from the beginning that nonprice signals and nonmarket control mechanisms play a very important role in the socialist economy. Yet recently it has been increasingly recognized that the so-called "quantity" adjustment processes in fact play an important role in *every* economic system. These take place without prices or at fixed prices (either way, without the equilibrating and incentive effects of permanently changing and adjusting prices). Part I hopes to contribute – by generalization of the experience of the socialist economy – to a better understanding of the nature of "quantity" adjustment.

[1]As opposed to the firm, the household is always – that is, even in the traditional management system of the socialist economy – highly affected by price, wage, and money. That is also why we have to leave its analysis to Part II of the book.

CHAPTER 2

The producer: Instantaneous adjustment

2.1. Introduction

We begin the discussion at the level of total generality. In sections 2.1 and 2.2 it will be left an open question as to whether we are talking about a socialist or a capitalist firm. This is because a few concepts must be clarified which will help in the remainder of the chapter and of the whole book in analyzing precisely the *differences* between the behavior of firms functioning within different systems.

In the modern large firm the function of production, and those of input purchase and output selling, are also separated organizationally. The former is fulfilled by the producer units or workshops under the direction of foremen, engineers, and production managers. The latter functions belong to the sphere of activity of the purchasing and sales departments. It may happen, particularly in small works, that these functions become intertwined or perhaps overlap. Yet on the plane of theoretical analysis we shall certainly wish to separate them. Chapters 2 and 3 have as their subject the firm as *producer*; in the subsequent chapters the firm will appear first as *buyer* then as *seller*. The producer carries out a *physical transformation*: he produces real output from real input. On the other hand the buyer or seller has the task of ensuring that the physical product finds a new owner, i.e. of performing a *transaction*. All three functions participate in the adjustment processes of the economy; however, they will be treated separately here. Meanwhile, of course, interdependencies between these roles will be referred to repeatedly.

2.2. Degrees of the producer's adjustment

Let us now begin to analyze the functioning of the firm as producer. Three degrees of producer's adaptation are distinguished.

Instantaneous adjustments are the responses of continuous adaptation of the firm to instantaneously prevailing circumstances. The production plan of the firm is given, either as a directive from higher authorities, or as the result of its own deliberations. The input–output combination to be applied is also given, and should be adhered to. Resource constraints are given: not only the fixed capital stock with a more or less permanent character, but also the stock of products, services, labor, etc. to be utilized as current input. If all these were given, what are the alternative adjustment targets and possibilities available to the producer? The answer to this question will be the subject of Chapter 2.

Short-term adjustment is the adaptation of the producer's plan for the next few months to the expected circumstances. This is a short-term phenomenon in the *temporal* sense: validity of the plan in question is short (e.g. three months), and the period between the preparation of the plan and the beginning of its fulfillment is also short (a few weeks). At the same time the expression "short-term" is also used here in the *Marshallian* sense: the adjustment in question is one that takes place with *given fixed capital*. It has three main components, interrelated but separable on the theoretical plane.

(a) *Determination of the aggregate production plan*. Here the question is: what will be the relation between the total resources available to the firm, and the total output it produces?

(b) *Determination of the input combination*.

(c) *Determination of the output combination*.

We deal with the short-term adjustment of production in Chapter 3.

Long-term adjustment is the adaptation of the plan for the next few years to the expected circumstances. In the *temporal* sense it is connected with the elaboration of five-year plans. (The book does not treat at all planning for a *very long* term, i.e. for 15–20 years, or an even longer period.) And again in the *Marshallian* interpretation the expression "long-term" also indicates that this does not take place with a given fixed capital but allows for its transformation. It is mainly connected to the allocation of *investments*, and will be treated in more detail in Chapters 9 and 10.

Short- and long-term adjustment may be interpreted as *learning*, using the word in the sense employed in the general theory of adaptive processes.

The producer learns lessons from the troubles and losses incurred in the course of instantaneous adjustment and will adapt himself to permanent difficulties with more fundamental changes.

2.3. Constraints on production increase

Now let us start our examination of the actual subject of the present chapter: the instantaneous adjustment of the producer firm. It is assumed that the firm is interested in increasing production. For the time being we will not dwell on its *motives*. It may be that it tries to increase its production as a result of a plan directive or a taut output target received from its superiors, or at least superior authorities expect it to make efforts at increasing production. It may be, however, that the firm acts in this way without any superior directives or expectation, only of its own volition, because it is driven in that direction by the desire to increase profit or by the urging of unsatisfied customers queuing for its products.

The question is, what *constraints* are met in the course of increasing production? Let us imagine a linear programming model describing the instantaneous production of the firm. What constraints would figure in the model? Since we do not want to make computations but only to demonstrate the train of thought, we need not be disturbed by the size of the simultaneous equation system. We do not aggregate at all: there are as many constraints in the mental experiment as exist in practice.

Three main groups of constraints can be distinguished.

(1) *Resource constraints*. Their general form is the following:

$$\boxed{\begin{array}{l}\text{The } i\text{th real input of}\\ \text{all producing activities}\end{array}} \leqq \boxed{\begin{array}{l}\text{instantaneously available}\\ \text{quantity of the } i\text{th}\\ \text{real input}\end{array}} \qquad (2.1)$$

These are *physical* constraints. Such are the material-, semifinished product- and part stocks instantaneously available to the firm, as well as workers of certain qualifications and of other particular abilities who are present instantaneously, functioning machines and equipment suitable for carrying out certain operations, etc. These – and only these – are the physical resources that can be used for production.

At this point our approach is to include resource constraints on production not just on the micro-level, but on the *submicro-level*. We penetrate down to the *elementary* production events taking place in workshops at each moment. It is quite possible that, considering the whole of a big enterprise, tens or hundreds of thousands of submicro-level resource constraints do exist. And when we think of the whole of the national economy, their number is in the millions.

(2) *Demand constraints.* Their general form is as follows:

$$\boxed{\begin{array}{l}\text{sales of the}\\ j\text{th product}\end{array}} \leqq \boxed{\begin{array}{l}\text{buyers' demand for}\\ \text{the } j\text{th product at}\\ \text{given prices}\end{array}} \tag{2.2}$$

Since we are now examining instantaneous adjustment, prices (if they affect the buyer at all) can be considered as given. Given that, demands are also considered as given.

The demand constraint exerts its effect on the producing workshop only in an indirect way. It is usually the firm's sales department that is in contact with the buyers. (This will be discussed in subsequent chapters.) Employees of the firm in charge of sales forward the buyers' demand to the directors of the firm, or perhaps directly to the production managers. In any case the producing workshops will perceive, relying upon instructions received from the directors of the firm or upon information transmitted by the sales department, whether they should increase manufacture of the *j*th product, or reduce it, or perhaps stop it entirely. In this way, under definite conditions, demand may restrict fulfillment of intentions to increase production.

(3) *Budget constraints.* Their general form is the following:

$$\boxed{\begin{array}{l}\text{expenditures of the}\\ \text{firm}\end{array}} \leqq \boxed{\begin{array}{l}\text{money stock + proceeds}\\ \text{of the firm}\end{array}} \tag{2.3}$$

We have only given a rough description of the budget constraint.[1] Thus, for example, items connected with credit have not been distinguished either on the left-hand or on the right-hand side. We shall revert to these

[1]Budget constraints exist for the firm as well as for the nonprofit institution and the household. Thus, the word "budget" is used here in a wide sense, i.e. as a synonym for the plan of financial expenditures.

If we want to talk expressly about the *state* budget (i.e. the financial plan of the state which is submitted by the Minister of Finance to Parliament for its approval), we shall always use the qualifier "state".

and, in general, to a more detailed examination of the composition and role of the budget constraint in Chapter 13.

Under certain institutional conditions (e.g. in a socialist economy) not just one but several budget constraints exist. Separately "labelled" limits may be put on expenses to be spent on wages or on investments or on imports. If we do not mention a distinguishing attribute, we shall include under the heading, "budget constraint", the *total* of expenditures and the *total* of money available. In other words, the constraint (2.3) formulates in a practical way the principle of "independent accounting of the firm": the firm has to cover its expenditures from its own proceeds.

The budget constraint, if it affects production, does so in an indirect way. It can prevent the firm buying physical resources: purchasing material and machines, employing workers.

2.4. Effectivity and hardness of constraints

Now only an objective function should be added to the constraints to give us a standard linear programming problem; we could immediately compute the optimum program of the firm. Yet we have not designed this model for advisory purposes; it is not intended to be an object-lesson for teaching operations research. This simple model of a firm is designed to serve as a framework for the *description* of the situation and behavior of the firm. We do not even pose the question of what the firm *should* do, or how it should choose *ex ante* the most expedient combination of activities. Our aim is to describe by this model *ex post* the actual functioning of the firm.

The following terms from the terminology of mathematical programming will be borrowed.

For some constraints, given in the form of inequalities, equality holds in the solution. Production fully utilizes one or the other resource; sales may reach the limit of demand; expenses may exhaust the available financial resources. The constraint is *effective* because in fact it restricts the selected activities. Production would have been larger if it had not hit effective constraints. It can also be said that the effective constraint is in fact *binding*. For other constraints, however, inequality holds (they are "not exhausted") in the solution of the programming problem. They are *non-effective* from the point of view of the instantaneous solution. It is as if they were not there at all, they do not influence the choice, they are "redundant", that is they *do not bind* activities.

It is always the comparatively narrow constraints that are effective; it is they that restrict efforts at increasing production. Comparatively broad constraints are not effective.

Other distinctions will be needed, too. Resource constraints are of a *physical* nature. They express the trivial truth that it is impossible to make something out of nothing. It is possible to apply, instead of the first input–output combination, the second or third one; but some kind of combination of inputs will usually be needed, and this is constrained by the available quantities of resources. Therefore the resource constraints cannot be exceeded: they are hard as rock.

The situation is different with the demand and the budget constraints: these express not a physical necessity but a *behavioral* regularity. These are laid down by people, and people can transgress them. A program within the constraints is satisfactory for the decision-maker, while he considers any overstepping unacceptable.[2] This type of constraint may also be called an *acceptance constraint*. The output stock, waiting for sale, has, in normal circumstances, a tolerance limit. Yet if it is exceeded, then it is exceeded. Transgression of the budget constraint means insolvency: this can also happen. *It depends on concrete circumstances, i.e. on social relations enforcing observation of the behavioral rule, how hard or soft the constraint is.* The hardness of the behavioral constraint has a graded *scale*: it can be almost as hard as the physical constraint, or of a medium hardness, or it can be expressly soft, that is it can be violated without trouble or consequence. A hard behavioral constraint may be effective, but it is not necessarily so – it depends on whether other constraints are relatively narrow or broad. On the other hand, a soft behavioral constraint (disregarding certain exceptional cases) can never be effective.

Now we have to hand the tools which should aid us in describing the situation and behavior of the firm from our point of view.

2.5. Resource-constrained versus demand-constrained systems

The constraint most often hit in the production of the firm, that is, which of the above discussed three types of constraints is effective, is deeply characteristic of the functioning of an economic system. First two "pure" types will be contrasted with each other. One is the "classical" capitalist firm. We are in the era before regular state intervention, that is, before

[2]This corresponds to the decision behavior that was called "satisficing" by Simon (1955, 1959).

Keynesian economic policy. We will disregard the peak of the upswing and will center our attention rather on the other phases of the cycle. The other "pure" type is *the firm functioning in the traditional socialist economic management system* (in the following called the *traditional socialist firm* for short). Its activities are controlled by detailed central instructions; it lives in an atmosphere of growth at a forced rate. The most important comparisons between these two pure types are summarized in table 2.1.

The decisive difference is already revealed in the first two lines. *With the classical capitalist firm it is usually the demand constraint that is binding, while with the traditional socialist firm it is the resource constraint.* With abbreviated expression (and to a certain extent simplifying reality) the following contrast can be formulated: the functioning of the classical capitalist firm is basically *demand-constrained,* and that of the traditional socialist firm is basically *resource-constrained.* This proposition plays a central role in the whole train of thought of my book.

Care must be taken in order to describe the situation exactly. Three qualifications must be emphasized.

(1) I do not claim that with the classical capitalist firm some of the resource constraints can never be effective – sometimes they can, but not often. And I do not claim either that with the traditional socialist firm some of the demand constraints can never be effective – the same can be said here: they can be effective sometimes, but not often. The statements

Table 2.1
The two pure types of firm, and types of constraint.

Type of constraint	Classical capitalist firm	Traditional socialist firm
Resource constraint	rarely effective	nearly always effective, more restrictive than demand constraints
Demand constraint	nearly always effective, more restrictive than resource constraint	rarely effective
Budget constraint	hard	soft
Production plan	autonomous: the firm lays it down at the level of demand constraints; within resource constraints	directive: prescribed by superior authority at the level of resource constraints; within demand constraints

are of a *stochastic character*: the probability of either one or the other type of event is prevalent with one or the other type of firm.

(2) As mentioned earlier, if a constraint is not effective, it is redundant and can well be omitted from the simultaneous equations. This holds here, too, for instantaneous adjustment, that is for daily or hourly production decisions. The present chapter is concerned only with this. Yet we must be aware that, indirectly and maybe with lags, noneffective demand constraints also affect the course of production. (This will be explained later, when discussing the short- and long-term adjustment of production and trade.) That is to say, the second row in the second column of table 2.1 does not mean, for example, that production of the traditional socialist firm is perfectly independent of demand. It means only that at the next moment of production, when the production plan is already given, actions are not constrained by the buying disposition of the customer, nor by the size of output stock accumulation that the firm management and the superior direction are prepared to tolerate with a given knowledge of demand, but primarily and above all by the inputs available.

(3) All that is said here is valid for the classical, traditional cases. The position of today's capitalist firm differs greatly from its classical predecessor; this will be treated briefly later in the book. And, the life of today's Hungarian firm is not exactly the same, either, as it was, say, 15 or 18 years ago. This will be discussed in more detail later on. The apparatus outlined here enables us to describe the historical changes that have been taking place in the position of the firm. However, for the time being let us content ourselves with analyzing the pure cases.

Let us revert to table 2.1. There is a close relationship between the third and the first two rows. The classical capitalist firm has a hard budget constraint. If it is insolvent, it will sooner or later become bankrupt. It can be granted a credit at best in advance of its future proceeds, which it has to pay back later with interest. It can buy only as much input as it can pay for from selling its products. Therefore it cannot produce more than it can expect to sell. It decides its production plans voluntarily at the level of demand constraints. ("At the level..." means that it produces approximately that volume; it might allow for some growth of its output stocks but, in the final account, it cannot depart much from expected selling possibilities.)

As opposed to this, the budget constraint of the traditional socialist firm is soft. If it works with a loss, that does not yet lead to real bankruptcy, i.e. ceasing operation. The firm is helped out somehow: it receives additional credit, or its tax is reduced, or it is granted a subsidy, or the selling price is raised – and, finally, it survives financial difficulties. Accordingly, its demand is hardly constrained by solvency considerations. The firm, as *buyer*, tries to acquire as much input as possible in order that shortage should not hinder production. The other side of the same phenomenon is that the firm, as *seller*, faces an almost insatiable demand. At least that is the situation with firms whose buyers are themselves firms: demand from such buyers is almost impossible to satiate. This insatiable demand "pumps out" the product from the firm. What is more, the superior authority determining the plan would also like to encourage the firm towards the largest possible production. The final result is that the production plan of the traditional socialist firm is set at the level of resource constraints. "At the level of resource constraints" does not mean that every resource is always fully utilized. We only mean by this that with the given composition of available resources, in consideration of the existing bottlenecks and with the given managerial abilities and organization, no more can be done. In any case, this planned level of production usually remains below what buyers would be ready to accept.

We have tried to summarize here in advance a long and complicated train of thought in a nutshell. Later, studying the details more profoundly, we shall revert on several occasions to these relationships. However, before continuing with our description of production, we shall refer briefly to preliminaries in the *history of economic thought* relating to the comparison made in table 2.1.

As regards the classical capitalist firm, the phenomenon in question is one that had an outstanding role in the thinking of Marx.[3] Marx's analysis of the deep contradiction which exists in capitalism between the tendency for unlimited enlargement of production and the limited absorption capacity of the market is well known.

Keynes's attention was centered on this sphere of problems.[4] He studied how effective demand for goods could be increased, in order that resource utilization, primarily the employment of manpower, should improve.

The contrast was very much emphasized in the Soviet economic discussions of the 1920s. Kritsman drew the following comparison in his study written in 1925: "in the commodity-capitalist economy there is a general *slack*, and in the proletarian-natural economy a general *shortage*".[5] In 1926

[3]See Marx [1867–1894b] *The Capital*. (For example, vol. III, ch. 15, section II about the conflict between the expansion of production and realization.)

[4]See Keynes (1936a, b) *The General Theory of Employment, Interest and Money*.

[5]See Kritsman [1925, 1929]. Kritsman's and Novozhilov's works were drawn to my attention by László Szamuely and Tamás Bauer.

Novozhilov contrasted "general overproduction" with "general shortage".[6]

Kalecki judged an essential difference between capitalist and socialist growth to be that the utilization parameters of the former are determined by the demand side, while those of the latter by the supply side.[7] Starting from Kalecki's thoughts the Czechoslovakian economists Goldmann and Kouba pointed out that socialist growth is constrained by various barriers: the man-power barrier, the foreign trade barrier, and so on.[8]

The comparison was also drawn for developing countries. For example, Hirschman differentiates between two types of growth. In one forward linkages are dominating: the producer pushes the users before him (this is analogous with our demand-constrained case); in the other type backward linkages are dominating: the user pulls the producers after him (as in our resource-constrained case).[9]

The message my book wants to convey is closely related to these thoughts. I try to make a step forward in that I shall endeavor to provide a more complete description and a more detailed analysis of the resource-constrained system.

2.6. Shortage and slack in production

In sections 2.2–2.5 we first described a general model of the firm then went on to draw a comparison between systems. In the remaining part of the chapter, however, with the exception of a few references, we deal only with the *socialist* firm that functions within the *traditional* socialist economic management system. We center our attention on resource constraints. Since we shall introduce some concepts which will be used frequently in subsequent parts of the book, we shall describe the resource constraint (2.1) once more, but now in mathematical form. For that we must supply some notation.

We examine the production of a workshop – let us say the cutting workshop of a screw factory. Let us consider the tth workday as one "instant" of production. Let us assume, for the time being, exclusively for the sake of simplification, that the workshop manufactures a single kind of product: a certain type of screw whose quantity is measurable by the piece. This output can, however, be produced by several, let us say by h,

[6] See Novozhilov (1926).
[7] See Kalecki (1970, 1972).
[8] See Goldmann-Kouba (1969, 1970).
[9] See Hirschman (1958).

alternative *technologies*.* Let us denote by $x_j(t)$ the volume of output manufactured by the jth technology on the tth day.

In the workshop a number, k, of different resources (labor ability of workers of different qualifications, a wide range of materials, machines, etc.) may be utilized. Technologies differ from each other in their requirements for labor of workers of different qualification, and for different kinds of materials and machines. With each technology is an associated set of fixed input coefficients; thus, for example, with the jth technology is associated the coefficients $a_{1j}, a_{2j}, \ldots, a_{kj}$. Some are positive, others zero – depending on the actual features of the technology. Thus, the technology determines the fixed *input combination* required for the production of a unit of the given output. Between the inputs of each technology there is, therefore, strict *complementarity*.

Let us denote by $r_i(t)$ the quantity of the ith resource available on the tth workday.

Now we can reformulate the resource constraint:

$$\sum_{j=1}^{h} a_{ij} x_j(t) \leqq r_i(t), \qquad i=1,\ldots,k. \tag{2.4}$$

production's requirement for the ith input on the tth day aggregated for all technologies	quantity of the ith resource available on the tth day

Let us assume that at the beginning of the workday the workshop manager has available the *plan* for that day. This was prescribed by the directors of the firm, or was determined by the workshop manager himself in the light of his knowledge of monthly or quarterly targets and agreed with the management as well as with other divisions of the firm.[10] The daily plan prescribes not only the quantity of production but also its technology. Let us give serial number 1 to the *technology specified in the plan*, while technologies 2, 3, ... represent input combinations deviating from the plan. Let us denote the plan by the symbol $x_1^{\text{plan}}(t)$. The workshop would like, according to our assumptions, to overfulfill the plan. In other words, it strives to make $x_1(t)$ as large as possible.

*The term *technology* used here and subsequently is more usually expressed in Western literature by the term, technique. Then a technology would be a collection of techniques. However, this wider use of technology is rarely needed in the present volume, so no confusion should arise. (*Editor's note.*)

[10] Since we are now dealing with the *traditional* socialist economic management system, we can assume that the daily plan of the workshop contains the targets – imposed on the workshop – from the production plan prescribed as the central plan directive for the firm.

Yet in its intention to augment the quantity $x_1(t)$ it repeatedly hits one of the resource constraints. Now this or that material is missing, now some indispensable machine breaks down, now a worker whose activity is needed has not come to work.[11]

It is assumed that the workshop management is incapable of flexible adjustment but it rigidly insists on the technology specified in the plan. (This assumption will soon be relaxed.) Thus, the rigid complementarity between the inputs required by the first technology asserts itself. Under such circumstances it is the relatively scarcest resource that delimits production. This is called the instantaneous production *bottleneck*. The maximum volume of production instantaneously attainable by the technology specified in the plan is denoted by $\hat{x}_1(t)$:

$$\underbrace{\hat{x}_1(t)}_{\boxed{\text{maximum volume of production by the technology specified in the plan}}} = \underbrace{\min \frac{r_i(t)}{a_{i1}}}_{\boxed{\text{maximum production allowed by the bottleneck}}}. \tag{2.5}$$

Now the conclusion to be drawn from this analysis can be summarized. *In a resource-constrained economy the first basic form of instantaneous adjustment to shortage is that the volume of production is adjusted to the current bottlenecks, i.e. to the instantaneously scarcest resources.*

Let us denote by $z_i^{\text{plan}}(t)$, the *shortage* of the ith resource, when workshop management insists on using the technology specified in the plan:

$$\underbrace{z_i^{\text{plan}}(t)}_{\boxed{\text{shortage indicator}}} = \begin{cases} a_{i1}x_1^{\text{plan}}(t) - r_i(t), & \text{if } \underbrace{a_{i1}x_1^{\text{plan}}(t)}_{\boxed{\text{input requirement according to plan}}} > \underbrace{r_i(t)}_{\boxed{\text{instantaneous amount of resource}}} \\ 0, & \text{if } a_{i1}x_1^{\text{plan}}(t) \leqq r_i(t) \end{cases} \tag{2.6}$$

[11]In Hungary the following joke serves to characterize factory life: On dying, an old capitalist goes to St. Peter who judges that he should burn in hell but, considering also the mitigating circumstances, he allows for one relief: the damned can make his choice between the capitalist and the socialist hell. Surprisingly, he chooses the latter. "You have been a capitalist all your life and now you want to go to the socialist hell?", St. Peter asks. "Yes", replies the old man, "because I trust that there, if they have a kettle to boil me in, there won't be water; if they have water, there won't be wood to heat it with; and if they even have wood, the devil in charge of lighting the fire won't arrive."

In discussions about shortage the question always arises: *in comparison with what* alternative should the shortage be measured? The *general* answer is the following: *shortage means that inputs required for the fulfillment of some serious intention were not available.* And now we have come to the first specification of the general answer. In this relation we consider the *daily operative plan* of the workshop as the producer's "serious intention", including also the volume and technology of production specified in the plan. We measure shortage in relation to the inputs required for the fulfillment of that intention.

Shortage cannot be described in terms of only a single scalar indicator. There exist several kinds of useful bases for comparison against which a shortage may be revealed. In formula (2.6) we gave one of these many bases of comparison and later we shall give others.

Now let us see the other aspect of the problem. We shall term every resource that is instantaneously unutilized, *slack*. The concept will be used as a broad, collective term. The word "surplus" has a pejorative tone: it indicates waste or loss. The word "reserve" has, on the other hand, an approving tone: a prudent manager accumulates reserves. Slack means dispensable surplus as well as reserve, it has neither a pejorative nor an approving tone. If a slack accumulates somewhere, it may be "good" or "bad", depending on the objectives of the decision-maker and on the circumstances. A long-unused machine may sometime replace one unexpectedly broken down, so that it passes from "bad" slack to "good" slack. And, conversely, material or labor reserves may later turn out to have been excessive. With better organization smaller reserves would have been enough.

Internal and *external* slacks are distinguished. The distinction is always relative; it cannot be interpreted in relation to the whole of the national economy, only in relation to the actual economic unit under examination. In our example the worker regularly employed by the enterprise and ready for work at his place, but without a task because of an instantaneous lack of material, represents an *internal* slack (unemployed on the job). As opposed to this, the unemployed queuing up for a job outside the capitalist factory is to be considered an *external* slack from the point of view of this firm.

As with shortage, we must now ask the question in relation to slack: *in comparison with what* alternative has the resource remained unused? As a first approximation we maintain the earlier assumption, namely that workshop managers insist on the technology specified in the plan. As was expressed in formula (2.5), in that case production hits the tightest resource constraint and that is why the volume of production for that day is $\hat{x}_1(t)$.

Let us denote by $\hat{q}_i(t)$ the internal slack of the ith resource on the tth day:

$$\underbrace{\hat{q}_i(t)}_{\boxed{\substack{\text{Internal}\\ \text{slack of}\\ \text{the } i\text{th}\\ \text{resource}}}} = \underbrace{r_i(t)}_{\boxed{\substack{\text{available}\\ \text{quantity}\\ \text{of the}\\ i\text{th resource}}}} - \underbrace{a_{i1}\hat{x}_1(t)}_{\boxed{\substack{i\text{th input required}\\ \text{by the maximum}\\ \text{quantity allowed by}\\ \text{the bottleneck}}}}. \qquad (2.7)$$

Slack is zero with a bottleneck and it is positive with the other resources. As we can see, in formula (2.6) quantifying shortage, the *planned* resource requirement is compared to the actually available quantity of the resource. On the other hand, in formula (2.7) quantifying slack it is the resource requirement of the *actual* maximum production that is opposed to the actually available quantity of the resource.

Shortage and slack are illustrated in fig. 2.1. With each resource is associated a horizontal strip; five strips are shown in our figure. Both the available quantities of resources and their actual uses are given in percentage form. Fulfillment of the production plan with the technology specified in the plan means 100 percent. This is represented by the heavy vertical line on the right-hand side of the figure.

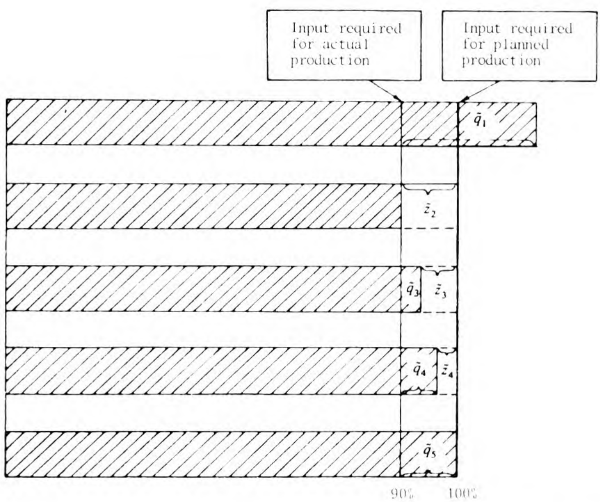


Figure 2.1. Shortage and slack in resource-constrained production.

The thin vertical line a little to the left of the 100 percent line represents actual production; according to this, the plan has been fulfilled only to 90 percent.

The shaded parts of each strip indicate the available quantities of the resources.

Let us first consider shortage. Of resources 1 and 5 enough was available for fulfillment of the plan. There was some shortage, however, of the other resources. This is indicated by the unshaded areas $\tilde{z}_2, \tilde{z}_3$, and $\tilde{z}_4$.

As regards slack, it was resource 2 that proved to be the bottleneck: there was no slack of it. On the other hand, slacks $\tilde{q}_1, \tilde{q}_3, \tilde{q}_4$, and $\tilde{q}_5$ are present for the other resources.

It is clearly demonstrated in fig. 2.1 that *as a consequence of complementarity, shortage and slack appear simultaneously in the same workshop.*

In fig. 2.1 we showed what happens if fulfillment of the operative plan of the workshop hits a resource constraint. This often happens in a resource-constrained economy and rarely in the demand-constrained economy. *The characteristic state of the demand-constrained economy is that there is slack of every resource.* As was explained earlier, here it is not physical resources but demand that is the effective constraint on the enlargement of production. For comparison we have drawn up fig. 2.2 which demonstrates the position of the workshop functioning in the demand-constrained system. If the volume of production could be raised to the limit allowed by resource

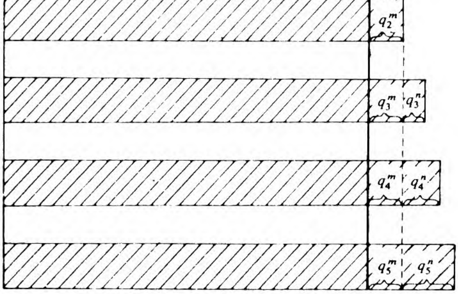


Figure 2.2. Slack in demand-constrained production.

2, production would be 5 percent higher than planned. Yet this is not allowed by the daily workshop plan adjusted in line with the demand constraint.[12]

Figure 2.2 divides slack into two parts. On the left-hand side the immediately *mobilizable* slack is to be seen. The slack q_1^m of resource 1 could be used immediately, because slacks q_2^m, q_3^m, q_4^m, and q_5^m of the complementary inputs, also immediately mobilizable, are also at hand. On the right-hand side q_1^n, q_3^n, q_4^n, and q_5^n are *nonmobilizable* slacks, since the required complementary resource 2 is not available.[13]

Now let us revert to fig. 2.1. There is no immediately mobilizable slack. The slacks of resources 1, 3, 4, and 5 cannot be put to use if complementary resource 2 is not available.

One more conclusion can be drawn from the two figures. It is not a matter of indifference whether the lengths of the various strips are *equal* or *unequal*. The more unequal they are, the more the exceptionally large slacks prove to be immobilizable, since their utilization is held back by bottlenecks; in other words, the more a considerable proportion of the slack must be regarded not as a "reserve" but as a real, unnecessary "surplus".

2.7. Forced substitution

Now we have explained what happens if fulfillment of the plan hits a resource constraint and the firm insists on the input combination specified in the plan. Yet the firm in question will not always be that rigid, but will try to adjust itself to shortage in some other way. Let us assume that the screw factory uses steel of quality 10 according to the technological standard for a certain type of screw. Now suppose this quality of steel ran out of stock and no new transport arrived. What is to be done? For example, the firm could decide to use quality 11 instead. It is a higher quality, therefore the material costs more. It may also happen that the extra cost cannot be collected from the buyer, in which case the loss would

[12]The fact that the workshop plan here is not derived from the central plan directive for the firm, but from the plan of the firm drawn up autonomously in consideration of sales possibilities, raises questions which do not belong in this section.

[13]In accordance with the subject of this chapter, we are only dealing here with *immediate* adjustment. If there is some time for adjustment, the momentarily lacking resource may, of course, be obtained in a short time. *Immediately* mobilizable internal slack may be replaced by *rapidly obtainable* external slack. This leads us, however, to trade among firms which we shall discuss in subsequent chapters.

be incurred at the screw factory. It may also be the case that the machine factory purchasing the screw is in such a difficult situation that it is forced to meet the extra cost. In the final account it is all the same. If the technology specified in the plan correctly determined the quality of the material, then from the social point of view utilization of the unnecessarily fine steel was certainly a loss. The reverse case may also happen: that steel of quality 9 is used instead of quality 10. The screw will be less durable. The buyer may not be told about it and he does not notice. Or he is told but is obliged to take it, because he is short of screws. Now there is a "saving" in material costs, but the *quality of output deteriorates. This is one of the most harmful effects of shortage.* In either case the firm deviated from the originally prescribed input combination. In the terminology of the model presented in section 2.6, it substituted another technology, for example number 2, 3... for the first technology.

In what follows this alteration of input combination as a result of shortage will be called *forced substitution*. It is clearly distinguished from *voluntary* substitution. If the firm switches over from the first technology to the second because it turns out that in this way an identical quality of output can be produced by less input or because relative prices have changed and the switching over saves costs or because the buyer wants a different material composition, then all this is voluntary substitution. But if the firm would not change voluntarily and is forced to make a revision solely because the necessary input is simply not available then we talk about forced substitution.

In a resource-constrained economy forced substitution is the second basic form of instantaneous adjustment to shortage. (The first basic form was discussed in the preceding section, i.e. technology and input combination are not changed, but the volume of production is reduced to the level allowed by the bottleneck.) It may be added that it appears that in a shortage economy forced substitution is the most frequent and most important form of instantaneous adjustment in production. The producer does not usually resign himself quietly to the fact that he has hit a resource constraint. He tries to do something. For example, an unskilled worker did not come to work. The producer may try to get the work done by a skilled worker. Or a machine broke down unexpectedly. Maybe the work can be done by hand? A part is missing that another factory ought to have supplied. Quickly the factory's maintenance workshop is given the task of trying to assemble the missing part somehow. ("Do it yourself" on an industrial scale.) A common characteristic of all the above-mentioned versions of forced substitution are *ad hoc* solutions and improvisations in order that the work should not stop and that a standstill in a small section should not spill over to other workshops of the firm.

Forced substitution is partly connected with *intertemporal* reallocations. For example, material of a certain quality was not available on Monday, but on Tuesday delivery is effected. Let us start immediately to work it up and, if we cannot finish during working hours, let us do overtime! That is to say, the material and normal working hours of Monday are substituted by the double quantity of material of Tuesday, the normal working hours and the overtime of Tuesday. An extreme form of intertemporal forced substitution is the rush-work at the end of a month or a quarter, when earlier missed production is made up for by overtime, Sunday work or extra night hours and even at the price of material waste and deteriorating quality.

As a result of forced substitution an impression may be formed that the shortage has been eliminated, or that it was not so bad at all. Production losses are smaller, some of the missing inputs are recovered from the slack. The absolute rigidity of complementarity, as described in section 2.6, will be somewhat softened. Nevertheless, the well-known phenomenon does not disappear: the shorter the time available for adjustment and for a change in technology, the stricter the complementarity that will prevail, and the less the opportunity for substitution that will emerge. Therefore in most cases, both shortage and slack will still appear together. Appropriate shortage and slack indicators may also be defined for the situation affected by forced substitution.[14] And, what is most important: *forced substitution itself is one of the most important indicators of shortage.*

2.8. Forced change in output composition

In the model of section 2.6 it was assumed, for the sake of simplicity, that the workshop produces a single kind of product. In reality, however, a workshop usually manufactures many kinds of products. Operative plans determine the composition not only of input but also of output.

If now there is a shortage of one or other of the inputs that would be needed for output to be produced in the original composition specified in the plan, the output composition might be modified. For example, material

[14]We shall disregard their formal definition. Their definition is identical with what is described in formulae (2.6) and (2.7), the only difference being that now the first technology is replaced by technologies 2, 3, ... describing the forced substitution input combination.

for screw type A, steel of quality 10, is not available. A decision may be made that screw type B will be produced, because steel of quality 9 is needed for that, and such steel is in stock. In many cases the producer does not manufacture what he *should*, but what he *can*, that is, what he is able to produce from the combination of available inputs. *In a resource-constrained economy the forced adjustment of output combination to available inputs represents the third basic form of instantaneous adjustment to shortage* (in addition to adjustment of production volume to bottlenecks and forced substitution among inputs).

In some of the cases only an *intertemporal* rearrangement takes place. Manufacture of screws of type B is brought forward and type A screws will be manufactured tomorrow when the material required for them is received. This is a comparatively fortunate case, as is the situation in which customers need A as well as B. In such instances the change of composition causes no trouble. It often happens, however, that the customer definitely wants A and yet the producer manufactures B. *That is how this form of instantaneous forced adjustment becomes one cause of an insufficient assortment of goods, not in harmony with customers' demands.* Here is the process leading to the appearance of so-called "shortage goods", that is of products in short and uncertain supply.

2.9. Joint emergence of the different forms of instantaneous forced adjustment

The different forms of forced adjustment may appear together within the workshop. For example, forced substitution among inputs occurs *and* the output composition is changed contrary to the plan. In other words, an input–output structure deviating from the plan, improvised under the pressure of inputs shortage, is applied.

But the various forms of forced adjustment may appear not only simultaneously, within the same workshop, but also in succession, linked together as in a chain, from one workshop to the next, or from one firm to the next. The rolling mill hit a bottleneck and did not deliver enough steel of quality 10 to the screw factory. (First basic form.) The screw factory manufactures, therefore, screws of type B instead of screws of type A. (Third basic form.) The machine tool factory which actually needs screws of type A is forced to fit screws of type B in the machine. (Second basic form.) *Shortage breeds shortage. The effect of forced adjustments caused by*

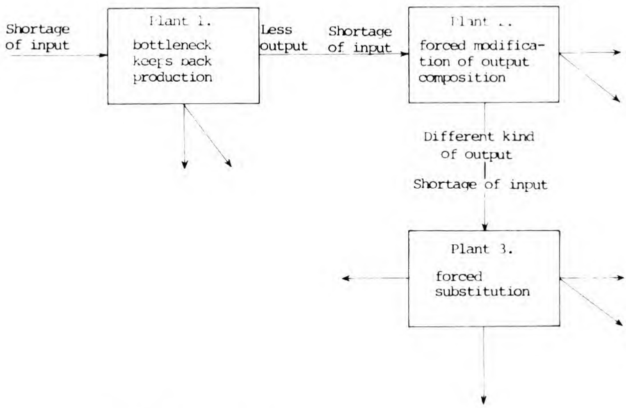


Figure 2.3. Spill-over of shortage phenomena in production.

shortage may multiply and spill-over in production.[15] It may retain or change its form in the course of transmission. The latter case, that is, the multiplication of shortage phenomena amid the metamorphoses of forced adjustments, is demonstrated in fig. 2.3.

There are countless possible variations in the linkages between the many different basic events. Some shortage phenomena do *not* spread further. Transmission may be stopped by the producer's output stock or the user's input stock which may accommodate an initial short-fall of production. Forced substitution may also localize the problem. In other cases, however, the above-mentioned multiplication and spill over takes place. One may say of "matter" that is not lost but only transformed. The same can be said of its "negative", i.e. *shortage* of material (and of other inputs): it is not usually lost but only transformed, either *inside* the factory or *outside* it, in linkages between factories.

The forced adjustments described in sections 2.6 – 2.9 require constant alertness of the production managers. Dispatchers charged with overcoming interruptions of production are to be found not just in workshops. In the traditional mechanism of economic management every leader in charge of a responsible post acts at least partly as a dispatcher: he intervenes in production matters by telephone as well as personally, he urges deliveries and arranges the regrouping of supplies and deliveries.

[15]Spill over is partly propagated by trade linkages between firms. We shall deal with them in more detail in Chapters 4 – 7.

Shortage not only has direct "material" consequences in the factory, but also, if the expression may be used, "psychological" consequences. The employee of the factory is not a computer producing a new input–output combination at each moment, but a living creature with sensitive nerves. The constant adjustment to shortage entails nervousness, confusion, and tension, whether there is a rigid retreat with stubborn maintenance of the planned technology, or a more flexible retreat with forced substitutions and rearrangements of output. It is accompanied by quarrelling and stress. And, the more frequent the hitting of a resource constraint, the more intensive the stress will become, and the more it saps time and mental energy from much more important tasks: the lasting improvement of technology and product quality.

Of course, it is not only in the traditional socialist economic management system that we can see permanent adjustments of production. In every system without exception production is continuously adjusted to momentary circumstances. The *system-specific question is*: *to what* is production adjusted? To physical bottlenecks? Or to changing requirements on the customer's part? Or to changes in the relative prices of inputs and outputs? We do not wish to put these questions as if we thought one adjustment criterion was "better" than the other. The customer may be capricious, and sometimes it does not matter if production does not follow his needs. It may entail a social loss if, sacrificing continuity of production, input–output combinations immediately follow every little movement in relative prices. At this point we state an observation without judgement: in a resource-constrained economy it is primarily shortage that forces the instantaneous adjustments of production to take place.

2.10. Observation and measurement

After giving an account in the preceding sections of the instantaneous adjustment of production to shortage we now go on to examine how these phenomena can be observed and measured. Two considerations lead us to deal thoroughly with the question of observation and measurement. (Not only here in this section, but repeatedly in subsequent chapters of the book.) One consideration is *epistemological*. The subject of my book is a *descriptive* theory of a shortage economy. I feel it extremely important to clarify whether the main concepts of the theoretical constructions developed in the book are *operational*. According to practice in the natural sciences the clear definition of a category is given in fact by indicating the method of its observation and measurement. From this point of view the important thing is not really to carry out observation and measurement immediately. What has to be discovered is whether the phenomenon reflected by the category in question is observ*able* and measur*able*. Our answer is affirmative: *observation and measurement of the phenomena described in this chapter may be difficult, but are not impossible.*

The other consideration that calls for a detailed discussion of the subject is the following: very important information is involved from the point of view of *practical economic decisions.* Central, medium- and low-level management authorities, as well as firms need to know, when, where, and of which products shortage has intensified or weakened, and how slack is developing. When the question is about information supplied to decision-makers, we cannot remain content, of course, with an examination of the theoretical possibility of observation and measurement, but we must also deal with the concrete practical conditions of their implementation.

In the following – not only in this chapter but all through the book – we shall use the expression *shortage indicator* as a broad collective term. We shall list under it all the measures that reflect shortage directly or indirectly. A characteristic example is the indicator defined in formula (2.6). A few additional shortage indicators will be suggested in subsequent chapters.

The expression *slack indicator* will also be used as a collective term. There is not only the indicator defined in formula (2.7), but any indicator reflecting the underutilization of a resource.

Shortage indicators will always be denoted by z and slack indicators by q. These symbols will be used if we talk in general (i.e. without concrete statistical definitions) about shortage and slack indicators. The symbol for the *ensemble of shortage indicators* is the vector $\mathbf{z}$, and that for the *ensemble of slack indicators* is the vector $\mathbf{q}$. To the basic symbols z and q some distinguishing mark will be added when denoting the various special shortage and slack indicators. Although, of course, each specific indicator must be defined separately, we shall apply the following conventions throughout.

Every shortage indicator is a non-negative variable. Its value is zero if the group of phenomena reflected by it is perfectly free of shortage. The value of the indicator reflects *shortage intensity*; the higher the shortage intensity in the group of phenomena described by the indicator, the larger the value assumed by the shortage indicator.

Similarly every slack indicator is also a non-negative variable. Its value is zero if the resource in question is fully utilized. The larger the underutilization of the resource, the larger the value assumed by the slack indicator.

After conceptual clarification and description of symbols we go on to explain a few *measurement principles* and *measurement tasks*.

(1) In the recording of shortage it is not enough to observe and measure *executed actions* – that is within the subject of this chapter – factory production and its inputs. The *intention* prior to action must also be described. Shortage means exactly that for the planned output and with the original technology specified in the plan insufficient resources were available.

The problem is that the official statistical records, everywhere in the world, mainly describe the processes of the real sphere: production, investment, trade, consumption, etc. Yet shortage is a phenomenon of the *control* sphere. The intentions of the decision-makers (or the different forms of intentions: aspiration level, plan, demand, claim for rationed goods, etc.) play an extremely important role in the control of the economy and in the regulation of institutions and organizations, that is, of the elements of the system.

The perception of "intentions" is not an easy task. (Although it is perhaps easier with production intentions than with buying intentions which will be discussed later.) It would not be right to provide some universal recipe. In some places it is possible and useful to rely on written plan documentation. In other places – instead of or as a supplement to this – the views of the firm's managers have to be discovered orally or in writing, either before or after the production activity we are concerned with.

(2) A task closely connected to the preceding point: let us state, at the submicro-level, what proved to be the effective constraint on increasing production. If not possible in any other way, it too can be clarified by questioning production managers: why do they not produce more? Is it because some of the physical resources are not available? Is there any mobilizable slack in production? If there is, what keeps them from utilizing it?

In this way we could arrive at the pre-eminently important measures of shortage and of slack: at the indicators expressing the frequency of hitting resource constraints and the size of mobilizable slack.

(3) For the measurement of shortage in production *all three* basic forms of instantaneous adjustment caused by shortage ought to be observed. That is, not only the first form already mentioned (the hitting of the

tightest resource constraint and restriction to the output limit set by the bottleneck) but also forced substitution and the forced modification of output composition.

None of this group of phenomena can be described by one composite indicator. Let us take, for example, forced substitution. I do not believe that all the losses caused by forced substitution could be summarized in a single synthetic indicator. Instead, a small group of indicators would be more useful. First of all the full scale of forced substitution in a plant can be calculated; for example, what percentage of production has been made by a technology different from the prescribed one? Besides, prevalence of the main types of forced substitution can be also measured (e.g. substitution of manpower in short supply by other manpower, etc.).

In accordance with our earlier interpretation, we shall include indicators expressing the forced modification of output composition in the list of shortage indicators (z indicators).

(4) It has turned out from previous sections of the Chapter as well as from points (2) and (3) that shortage as well as slack can only be described by the *ensemble of a large number of indicators. Neither shortage nor slack are scalar categories but vectoral ones.*

(5) Shortage and slack must be measured independently of each other. Let us think over the following problem.

Production is observed at the submicro-level in the most detailed breakdown possible and at each moment. In that case we can make a statement of the following kind about a single elementary event of production: on 17 May at 9.07.14 hours in workshop B the worker operating the third machine ought to have processed steel of quality 10. Either this material was in a pile before him or it was not: he could wait for it, or he could substitute another material for it. In other words: there was *either* slack *or* shortage of the material. Assuming a *given fixed intention* (steel of quality 10 *must* be processed) and "*absolute fineness*" of observation the two states are mutually exclusive. *Yet if at the micro-level the aggregate state of a bigger unit (several workshops, several firms, several inputs) is described, or all events over a longer period, or we aggregate in both senses (i.e. we describe a bigger unit for a longer period), shortage and slack will no longer be mutually exclusive states. Then it is possible and even likely that shortage and slack will exist side by side.*

Although the logic of this train of thought is easy to understand, this is not a self-evident truth. On the contrary, traditional economic thought considers shortage and slack to be mutually exclusive categories even at the micro- and macro-levels. We will return to this question on several occasions, since it has highly important implications both for theoretical analysis and for economic policy.

If the preceding thoughts are accepted, it becomes apparent what a grave mistake it would be to "net out" shortage and slack, or "minuses" and "pluses" in measuring aggregates. Let us take a simple example. In a workshop the staff required is 100 persons. On Monday and Wednesday 90 persons came to work, on Tuesday and Thursday 110 (workers were directed here from the neighboring workshop). It would be a complete evasion of the problem to supply statistics merely informing us that on average 100 workers were available. Instead, three items of data at least have to be recorded: the proportion of days of labor shortage and that of labor surplus; average shortage on days of labor shortage (10 percent) and average slack on other days (10 percent).

This is, of course, just a simple example to serve as an illustration. In actual measurement more complex indicators than this will often be needed. Here only the principle will be emphasized: statistics of shortage and of slack are needed separately, and they should not be "netted out".

(6) Shortage, slack, forced substitution, the modification of output composition, and other related processes appear at every moment in the form of *millions of elementary events at the submicro-level*. Of course, it cannot be expected from any observation and measurement that they should fully describe all of these. What can in fact be expected is a *statistical description* of shortage phenomena. As explained in paragraph (4), various kinds of indicators are used to represent the main types of events related to shortage and to slack, with all of which *probability distributions* could be associated. (Or at least some approximation to the distributions.)

The expression "statistical description" is used in a similar sense in which, for example, physicists talk about "statistical physics". They use this to describe the world of elementary micro-phenomena in a stochastic form of expression, without "jumping over" to the world of macro-physics of large aggregates. A similar thing is needed in the description of economic systems. The train of thought will be conducted further in the ensuing chapters of the book.

After surveying a few of the *principal* problems and general tasks of measurement we shall add a few remarks on the *practical* organization of observation and measurement.

With due inventiveness we can measure the intensity of shortage, the extent of instantaneous forced adjustments and the size of slack in quite a number of different ways. It must be recognized, however, that almost all measurements require labor- and cost-intensive observations. For example, standstills caused by input shortages would have to be recorded, as would the causes of the enforced idling (material shortage, current failure, the worker has not come to work, machine trouble, etc.). Or records must be kept of forced deviations from the planned technology and product composition. A few examples were mentioned above when discussing measurement principles.

The task of measurement is largely diminished, however, by the fact that complete observation is not necessary. It would be sufficient to observe a suitably selected *sample* of events, which could adequately represent a firm, a sector, or the whole economy. And, of course, no objection can be raised against *partial aggregations*: for example, the summation of one or other measurement over the whole of an industry or over the whole of an input group. It is not the task of the present book to submit proposals which go into the details. It seems certain, however, that the systematic observation of a few dozen indicators, and the description of their distribution, are sufficient to characterize the current economy-wide state and dynamics of shortage, slack, and forced adjustment in production quite adequately.

If such observations and measurements have not been carried out before, the cause does not lie in its impossibility or excessive cost. The only explanation is that their real importance has not been recognized, thus no decision has been taken to organize the recording. A long time passed before the systematic observation and measurement of unemployment was introduced in every capitalist country. In socialist countries the shortage problem has become topical. Sooner or later the systematic measurement of shortage indicators will be organized.

2.11. Normal shortage and normal slack in production

Let us assume that we have in hand the most characteristic summary indicators of shortage phenomena in production for a country. After a few years of observation their intertemporal averages could be determined.

I would like to set up six hypotheses in this connection.

(1) *The main shortage and slack indicators of the system show a certain degree of stability*. While the institutions and external conditions of the system are more or less stable, the intertemporal average of these indicators will also be quite firm. In what follows these intertemporal averages will be called the *normal values* of the indicators in question. I shall talk in

this sense about the *normal shortage and the normal slack* of the system. (Normal values will be marked by an asterisk.)

The question is not simply about the trivial point that these indicators – just as any other random variable – also have a mean value. My hypothesis is that in the system feedbacks and control mechanisms are functioning which drive shortage or slack deviating from the norm – either above or below it – back to their normal levels. In the latter parts of the book such mechanisms will be treated repeatedly.

(2) *In reality no system exists in which normal shortage or normal slack would take the extreme value of zero.* There is no "perfect" system, with zero shortage and zero underutilization of resources.

(3) While it is a common property of all systems that these indicators are definitely positive, it is already *deeply characteristic of individual concrete systems, what values z^* and $\boldsymbol{q}^*$ – the normal shortage and normal slack vectors – take.*

(4) *The normal value of shortage indicators in the traditional socialist firm is considerably higher than that of similar indicators of the classical capitalist firm.* According to this hypothesis the indicators which describe the constraints usually hit by expansion of production would show the resource constraint being hit most frequently in the case of the traditional socialist firm.

(5) We make no hypothesis concerning *total* (mobilizable plus non-mobilizable) slack. We can, however, hypothesize about the *immediately mobilizable* slack: it is very low in the production of the traditional socialist firm; much smaller than with the classical capitalist firm.

(6) Hypothesis (1) does not mean that a concrete system is once and for all tied to its own normal shortage and normal slack from which it cannot depart. *The norms*[16] (in the present context this means the normal values of shortage and of slack) *themselves are social formations which come about as a result of historical development and are made firm by social conventions.* Considerable changes in institutions, control mechanisms and, with these, also in social expectations and conventions may also entail modifications in norms.

[16]The word "norm" is used in two different interpretations. One interpretation is *average, not extraordinary*, not sick. It is with this interpretation in mind that medicine talks about normal body temperature or normal leukocyte-count. The other interpretation of a "norm" is what is *desirable* or expectable. It is with this interpretation that the term "normative value judgement" is usually spoken of. In this book I shall always use the word with the *first* interpretation in mind.

It seems, for example, that as a consequence of the Hungarian economic management reform and other economic policy changes the normal values of shortage and slack have been shifted. Several shortage indicators would today show lower values, that is nearer to zero (but of course, still above zero) in the second half of the 1970s, than was the normal value in the first half of the 1960s.

The six hypotheses mentioned above are supported by a logical consideration of the problems as well as by sporadic observation.[17] The observation and measurement methodology outlined earlier in this chapter will render *these hypotheses empirically testable*.

It is appropriate here to make a few remarks on terminology. I have mentioned the *normal values* of shortage and slack. I do not endeavor to define abstractly and in general terms the expression "normal state". We can perceive the meaning without an exact definition. There is a system whose external conditions as well as internal behavior exhibits regularities which are more or less stable. The "regular" values of the main state variables of this system can be called the "normal state". Small external or internal disturbances in the system make the values of state variables depart from their normal levels, yet even with the fluctuations, it is this "normal state" that expresses the prevalent tendency.

We must avoid associating any value judgement – expressed or implied – with the term "normal state". A precapitalist society may have been stagnating for centuries. Obviously, that is its normal state. Nobody would think of calling this state "good". To assert that a system is in its normal state is neither praise nor rebuke; it is not condemnation, but not an excuse either. By this, we say no more and no less than that a system functions in accordance with its own inner nature. It is an extremely important part of the scientific analysis of a system to understand what the normal values of the main state variables (or their normal path in time) are. The comparison of systems seeks to answer the question: how does the normal state of one system differ from that of another system?

[17]In the judgement of the firm's managers it is difficulties with material supply, and problems of cooperation with input-supplying firms that should be considered among the gravest sources of loss, of the factors that render it difficult to increase productivity. See Román (1973).

There is some valuable empirical material about the relationship between shortage and slack-formation in the studies of Chikán (1977) and Nemes (1976).

CHAPTER 3

The producer: Short-term adjustment

3.1. Introduction

In the previous chapter we examined, how the *continuous* adjustment of the firm to the circumstances prevailing at each *instant* takes place. Now we go on to examine *short-term* adjustment. As already explained in section 2.1, we understand thereby mainly the drawing up of the plan for the coming months (three months or a year). The planner tries to take into account the expected circumstances – potentialities of the firm and demand for the firm's output – and to adjust targets to them.

In passing from "instantaneous" to "short-term" we have in fact skipped some intermediate stages. A number of measures may be taken which cannot be implemented immediately but only after a certain reaction period; and yet there is no need to wait for their inclusion in the next quarterly or annual plan. For the sake of brevity we will not discuss these intermediate stages.

There is a close relationship between the subjects of Chapters 2 and 3, that is between instantaneous and short-term adjustment. The more successful the latter, the less improvisation or instantaneous forced adjustment is necessary. And, conversely, the frequency of instantaneous forced adjustments and their troublesome consequences may serve as a signal influencing the elaboration of the next short-term plan.

As in the previous chapter, our attention is again centered on the producer firm functioning in the *traditional* socialist economic management system. Also, we shall make brief references to experiences to the postreform situation in Hungary.

3.2. The aggregate production plan

Our first subject concerns the determination of the aggregate quarterly or annual production plan. Aggregation is done usually at the constant prices of some base period.

Of course, the plan for aggregate output and the detailed targets (output targets for a few priority products and input quotas) are established together. Yet now, on the plane of theoretical analysis, let us restrict attention to the problems of determining the aggregate output target. In this section the expression "production plan" will always indicate the aggregate target, even if this is not expressly stated.

For the traditional socialist firm, the plan determined by the superior authority is usually "taut". What does this tautness mean? Before giving an answer, let us pose two more questions. It was observed in the previous chapter that the firm hits resource constraints again and again. Does this not mean that the plans are unrealizable? How can this permanent hitting of resource constraints be reconciled with reports which always state: plans have been fulfilled?

Let us distinguish three different situations.

First situation: the firm has received a production plan and such resources along with it that it was able to fulfil and perhaps even overfulfil the output target without once hitting a resource constraint. And it could attain all this by the technology specified in the plan, without any forced substitution, or any improvised modification of the planned output combination. Superior authorities would later say of such a plan that it was *loose*.

Second situation: the firm has finally fulfilled or even overfulfilled its production plan. In doing so, however, it hit resource constraints. It adjusted itself to the situation in many different ways: it made forced substitutions repeatedly, including some intertemporal restructuring. If it lagged behind the daily plan on days when resources were scarce, it did extra work at another time and made up for it by a "rush". It deviated from the concrete composition of output centrally prescribed or planned within the firm. It manufactured a product of poorer quality or, if not of a poorer, at least of a different quality than was originally prescribed. And it did all this in order not to lag behind its aggregate production plan. Superior authorities would later say that this firm received a *taut but realizable* plan. In the traditional economic management system this is usually the aspiration level of superior authorities: they try to prescribe taut but realizable plans for the firm.

And, finally, the *third situation*: the firm effected various forms of instantaneous adjustment, but is nevertheless unable to fulfil the production plan. It may be that in such a case superior authorities would blame the firm. Within the firm, however, it will certainly be said that the production plan was *overtaut* in relation to the allotted resources.

Now consider the combined result of production by all the firms. While firms in the second and third situations were continuously hitting resource constraints, the third category even lagging behind the plan, this could be compensated by overfulfilment in the first and second categories. Overall, the aggregate plan of the sector or of the whole industry could be fulfilled and even overfulfilled.

In this way we have arrived at an interpretation of the "tautness" of the production plan. We do not assert that our description captures all aspects of the concept of "tautness", but we may have succeeded in grasping one of its most important features. Tautness of the plan is a stochastic category. The tauter the plan, the more probable it is that in the course of its implementation the firm will hit a resource constraint and that forced substitutions as well as forced deviations from the planned output composition will be effected. If firms receive tauter (or even overtaut) and less taut (or even loose) plans in an appropriate distribution, then fulfilment of the aggregate production plan of the sector or of the whole industry seems to be assured with high probability.

This may seem a highly complicated stochastic decision problem, yet it is regularly solved in the practice of production planning without any particular mathematical apparatus. A *control mechanism* is functioning that sets plans at the appropriate level of tautness. To facilitate an explanation we present fig. 3.1.

Let us assume that it has been decided what resource quotas the firm should receive, according to the plans, in the coming three months. Of

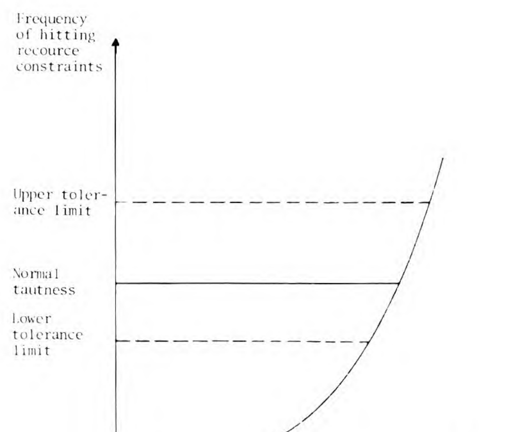


Figure 3.1 Tautness of the production plan.

course, the resources actually available may deviate randomly from this, and will often be more scarce than planned. The question is, what should the planned level of aggregate output be. This is represented by the horizontal axis.

Tautness of the plan can be expressed by various kinds of indicators; one possibility is shown on the vertical axis. How often will the firm hit a resource constraint in the course of the coming three months?

Up to a certain level of production the curve follows the line of the horizontal axis: the expected frequency of hitting a resource constraint is zero. Above that level it takes a positive value and is rising. The rising line becomes increasingly steep: the hitting of resource constraints becomes more frequent. (Of course, the exact shape of the curve has been chosen arbitrarily. This is, however, not important, since it serves only to illustrate a train of thought.)

Three horizontal lines are seen in fig. 3.1. The thick line in the middle is the *normal tautness* of the plan. Forced substitutions, modifications of output composition, and other improvised adjustments have a customary level which is not surprising either for shop managers or for workers. They would be more surprised if this were entirely eliminated and production were going on without any difficulty caused by shortage phenomena.

The lower dotted line is the *lower tolerance limit* of tautness. This is the acceptance constraint of the superior authorities determining the plan. If difficulties grew even less frequent than this, they would feel that the plan could be realized much too easily and without any effort: the plan is "too liberal"; it has insufficient motivating force.

The upper dotted line is the *upper tolerance limit* of tautness. This is the acceptance constraint of the firms. If that were overstepped the situation in the workshop would become unbearable. Problems caused by shortage would become so frequent that they would totally disrupt any acceptable order of production.

The mean value – normal tautness – and the two extreme values – the lower and upper tolerance limits – are historically developed magnitudes fixed by social conventions and practice. The mechanism regulating the tautness of the plan is built upon them. *Feedbacks* are functioning according to the following logic.

(1) If the hitting of resource constraints (or, in a more general form, problems caused by shortage and forced adjustment) is expected to become too frequent as compared to the norm, the plan must be lowered; in the reverse case it must be raised.

(2) If there is an imminent danger that the frequency of hitting resource constraints (and other troubles) will exceed the upper tolerance limit, the

plan must be determined at a lower level; if the lower tolerance limit is unlikely to be reached the plan must be revised upwards.

The first kind of feedback is called *control by norm* and the second *control by tolerance limits* (*or by critical values*). These two mechanisms appear in the control sphere of every economic system. Systems differ from each other, among other things, in *what* is controlled by these two mechanisms and, furthermore, in the nature of the norms and tolerance limits and *how they are formed*. Later we shall deal at length with similar control mechanisms. We shall also talk about their mathematical modelling. But here we go no further than their verbal description.

Associated with every feedback control there is a *signal system* which supplies information. In the present case a complex signal system is functioning in which many different types of information are flowing. Their common feature is that they are *nonprice signals*, i.e. "quantity" signals. There is no change in prices to convey information about shifts in relative scarcities, shortages, and slacks. An important signal is provided by observation of the fulfilment of previous production plans. The superior authority determining the plan observes this fulfilment. If the firm has repeatedly lagged behind these plans, it may indicate that the plans were overtaut, so that next time a less taut plan must be determined. If, however, the firm has overfulfilled the earlier plans, it may indicate that it had too easy a task. Let the plan therefore be tauter on the next occasion. This kind of feedback is well known to firms: they call it "planification"* of the attained higher level. This is what puts them on guard: although overfulfilment of the plan is usually awarded by bonuses, it is not wise to go too far in this direction.

Data on plan fulfillment are summed up in *written* reports. These are, however, supplemented by a further signal: the *spoken languages*, *the "voice"*.[1] Not even in the strictest periods of centralization was the plan directive a one-sided dictate. It was always preceded by some dialogue between those determining the plan and those implementing it. This is called "plan bargaining": those determining the plan would like the firm to produce more output from less input. The executor of the plan wishes the

*This word is a literal rendering of a Hungarian word which has no good English equivalent. It means that performance in the previous period is built into the plan for the current period. (*Editor's note.*)

[1]I borrow the terminology here from the essay by Hirschman (1970) "*Exit, Voice and Loyalty*". In this work it is shown extensively that protests, claims, forcefully presented opinions advanced orally by the people directly interested in the consequences of a certain process, i.e. "voice", can be an important and effective signal of equal rank with many other "silent" signals. (For example with the termination of participation, i.e. exit, or with the price signal.)

contrary: a lower output target, and with it a more generous input provision. In the bargaining the voice of the firm warns of the upper tolerance limit: a large upwards deviation from the norm; the voice of the superior authority warns of the lower tolerance limit: a large downwards deviation from the norm. It is in the dispute between the two different "voices" that a compromise is found which, so to say, "calibrates" – in regard to the totality of firms – the plan at the normal and acceptable level of tautness.[2]

The mechanism outlined above for making the plan taut guarantees that in the traditional socialist firm the aggregate production plan should be determined *at the level of resource constraints*. (See the last row of table 2.1 drawing a comparison between systems.) It is now even more clearly understood: the expression "at the level of resource constraints" does not mean that every firm utilizes every resource at 100 percent all the time. Not at all. This only means that nearly every firm frequently hits one or other resource constraint, and these bottlenecks bar the way to instantaneous increases in production. The mechanism of taut planning pushes the firm's production up to that level.

One more thing must be added to what has been said: *norms and tolerance limits playing important roles in control are historically determined social formations, themselves changing with the deeper changes of society*. In the period of socialist economy immediately following the revolution, shock-work and Sunday and holiday rush were regular. Yet today shop managers as well as workers expect a much smoother material supply and solid organization of production. If, therefore, a comparison were drawn between norms and tolerance limits of "tautness" in today's socialist economy and those of twenty or thirty years ago, it would be found that those of today are at a much lower level. This, however, does not prevent us, in a short-term analysis, from considering norms and tolerance limits as *given* for the control mechanism in existing social conditions.

So far, we have been discussing the traditional case, where the superior authority determines the plan for the firm. In Hungary, however, since the 1968 reform no central plan directives have been given to the firms indicating what they should produce in the next year or quarter; each firm is entitled to make its own decision about its short-term production plan. This is a change of extreme significance, which has a number of favorable

[2]For an empirical description of "plan bargaining" see the author's book *Overcentralization* (Kornai, 1957, 1959). Several attempts have been made at modelling "plan bargaining"; among them the game theory approach of Johansen (1977) is remarkable. See also Dolan (1976).

effects; yet they go beyond the scope of this book. At this point we shall only ask one question: as a consequence of the reform, has the resource-constrained character of production ceased? Is the plan, determined by the firm itself, no longer a taut one?

My hypothesis is that *tautness of the plan has remained, although at a somewhat less intensive degree. Production – with forced substitution and other forms of forced adjustments – rises to the level at which it hits resource constraints, that is bottlenecks caused by shortage.*

The hypothesis is *empirically testable*. Sporadic observation supports the statement. There is only a small number of cases where every complementary resource is available in a workshop with immediately mobilizable slacks of them all, and the workshop still refrains from increasing its production. It is rather exceptional for the producer to refrain from expanding production solely because the product is unsaleable. In other words, it is only in exceptional cases, and even then for only short transitory periods, that the demand constraint proves effective, that is really limits activity.

Without aiming for completeness, we mention three factors which can encourage the firm to produce up to the limits set by resource constraints.

(a) Shortage itself encourages it to do so. If buyers queue outside the door (symbolically), it is hard to resist the "pulling" effect of demand. The more so as the unsatisfied buyer has a strong "voice". He could protest to the superior authority of the firm or to political and social organs if it turned out that the firm could produce more yet does not do so.

Shortage is, as will also be seen in many other kinds of interrelation, a vicious circle. The firm as *seller* suffers the influence of the impatiently queuing buyer; it is obliged to push production up to constraints determined by bottlenecks. It develops a *feeling of want*: it feels that it is the comparatively scarce resources that are hindering the increase of production. Therefore, now as a *buyer*, it joins the queue of those demanding more input.

(b) Superior authorities cannot, in the legal sense, dictate what a firm's short-term production should be. On the other hand, they can informally "put pressure" on the firm. For example, the Ministry can invite directors of the firm and persuade them to produce more, in case a restrictive policy were needed. And it would certainly be done, since there is a very deeply rooted view that "hoarding" of resources is wasteful and a social loss.

(c) The firm also strives voluntarily for a quantitative augmentation of production. Later, it will be explained how the true interest of firms has developed since the introduction of profit-sharing and other financial incentives. Here just one thought will be anticipated. Today's Hungarian firm is in a situation where it is able sooner or later to pass on to the buyer or the state budget any increase in production costs. Given this, therefore, it is in its own interest (including the firm's profit motivation) to go to the limits of capacity, even if the initial increase in marginal costs is not promptly counterbalanced by an increase in marginal proceeds.

One final general remark. The group of phenomena connected with the tautness of the plan is called the *quantity attitude* or the *quantity drive* in socialist economic literature and everyday jargon. (The latter expression will be used in this book.) In order to clarify the phenomenon described in many economic books we shall only add one point: *its appearance cannot be explained solely by the bonus for managers related to overfulfilment of the plan. Deeper interior motivations develop it very powerfully.*

3.3. Input combination

Let us now turn to the planning of input combination. As was mentioned earlier, three degrees are distinguished.

Within *instantaneous* adjustment the firm continuously adjusts its input combination to the instantaneous available resources. Within *short-term* adjustment – and now this is going to be explained – the firm introduced a more permanent modification to its input combination, responding to apparently permanent resource shortages, without, however, changing its fixed capital stock. Within *long-term* adjustment the firm is concerned with capital formation and also modifies technology in this way, as a further reaction to permanent shifts in resource supply.

Let us take as an example the response of a firm to labor shortage. Instantaneous adjustment: on Monday transport worker X. Y. did not come to work. The foreman asks N. N. to replace him on overtime (forced substitution). Short-term adjustment: since the shortage of workers concerned with transporting work within the factory occurs frequently, the workshop is now reorganized so that workers operating the machines themselves pass the work-pieces from hand to hand for further processing. Long-term adjustment: material transport in the factory is mechanized.

Our hypothesis in regard to short-term adjustment is the following:

The lasting adjustment of technology to chronic shortage of one or other kind of resource takes place sooner or later, even though with lags, and merely in response to the repeated experience of hitting resource constraints. It also takes place even if relative input prices have not changed and the introduction of the new combination has not been preceded by a systematic search for a solution minimizing costs in money terms.

The hypothesis requires empirical testing. I feel certain that the hypothesis is true, in my cautious formulation. What I am uncertain about is the concretization of the meaning of the expression "sooner or later" in the above proposition. I am also uncertain about the question of which is quicker and more reliable: the control mechanism based on relative prices and cost minimization, or that based on the direct perception of relative shortages. It may be charged against the price mechanism that it functions with lags: it may take a long time before relative prices react to changing relative scarcities and it again takes some time before the input combination is adjusted to new relative prices. All this is true. Yet it can equally be said that it may also be a long time before the workshop manager or some Ministerial official realizes that the relative scarcity of resource A in relation to resource B is not a transitory but a lasting phenomenon. And, as regards action following perception, it may again take a long time before he brings himself to introduce more permanent changes in technology instead of improvised forced substitutions.

The question of motivations is also involved in the comparison. Profit incentives may play a part in technological adjustment based on relative prices. I do not now want to go into the analysis of how strong this is with socialist firms applying profit sharing; this will be the subject of a later chapter. What I wish to emphasize here is the following observation: even a workshop manager or a Ministerial official who has no direct material interest in profit is not indifferent to resource shortage. Uncertainty of material or manpower supply, difficulties in implementing forced substitution, the tension and hard-thinking preceding improvisations all encourage him to try to find a more durable solution. The struggle for his own tranquillity and smoother work for his colleagues and subordinates prompt him to develop an input combination better adjusted to the more permanent proportions of resource supply.

It is quite possible that it is in the area of these accompanying circumstances that we may find one of the advantages of the mechanism based on relative prices: it is more impersonal and more objective, and perhaps its operation involves less stress; it is presumably simpler and more reliable. Those who finally decide about technology are not compelled to compare the relative frequencies of hitting resource constraints, relying on memory and impressions; instead, he can carry out relatively simple cost calculations. In the comparison of mechanisms based on relative prices and on the direct perception of repeated shortages it has been assumed in both cases that technological modifications follow *posterior* signals. Relative prices have already changed – and then comes a new technology. Or, resource shortage has already been revealed – and then comes a new technology. In reality there are, or course, *anterior* signals as well. These terms are used to denote signals following and preceding some event or change of circumstances, respectively. The decision-maker intending to minimize costs tries to obtain information about future prices. And in the mechanism without price-signals plan targets relating to future resource availabilities can be taken into account in current decisions. From this point of view planning has vast potential advantages. It is another question just how far these possibilities are realized in practice by socialist firms. Experience shows that *incentives* are just as important as *information* about relative scarcities. Man is inclined to repeat his own routine, which also holds for the workshop manager. He will bring himself to make an important change if he runs head on into a resource constraint, and actually experiences the "shock" himself.

The question needs to be further examined both theoretically and empirically. To understand subsequent ideas in this book it is enough now to know that perception of the recurrent, chronic relative shortage provides a *sufficient signal*, and the wish to avoid the forced measures taken to accommodate to the consequences of shortage provides a *sufficient motive* for a short-term adjustment of the input combination.

3.4. Output combination

With regard to the short-term adjustment of the firm's output combination let us first of all differentiate between two ways of organizing production: the firm manufacturing *to order* and the one manufacturing *for stock*.[3] If the product has numerous individual characteristics, the firm starts production only to order. That is the situation, for example, in the shipbuilding industry, as well as with much of the large special factory equipment and machines. If, on the other hand, a product is manufactured in large lots, it is not necessary to know beforehand the customer's particular requirements, so the product can be manufactured for stock.

Of course, the distribution of the two types of organization does not only depend on the qualities of the product but also on the market situation. In the "seller's market" the buyer may also place his order in advance for

[3]See Belsley (1969) *Industry Production Behavior: The Order-Stock Distinction.* We return to the distinction in Chapter 5.

mass-produced goods, with a view to getting his turn with certainty and at the soonest possible moment. And, conversely, on the "buyer's market" the producer may not wait for an order but manufacture for stock even such products as he would in fact prefer to manufacture only to order.

In determining output combinations the producer receives important signals from the dynamics of output stocks and from the backlog of unfilled orders. It is worth changing the output combination in favor of product A and at the expense of product B

(a) if the output stock of product A has diminished while that of product B has grown; or

(b) if the output stock has been entirely exhausted and buyers are even queuing up for A, placing orders, while product B is still in stock; or

(c) if there are queues for both A and B, but there are more orders placed for A – the queue is longer – than there are for B.

Signals provided by stocks and by the backlog of unfilled orders can be completed or intensified by the "voice". The customer who feels pushed into the background may protest and demand more loudly, asking the superior authorities to intervene.

In the final account, a *feedback control mechanism* functions here which is similar to that treated in the preceding two sections. This feedback makes its effects felt even within the traditional forms of strong centralization. It is true that the superior authority of the firm gives highly detailed prescriptions for the output combination. Proportions of the combinations are rather rigid in some places; they change but slowly. If a change does take place, however, it is often in response to such a feedback. Output combination is modified in favor of product A and at the expense of product B because the queue has grown too long, and because there are too many and too strong complaints resulting from the shortage of product A.

Yet it is not only on the superior authority but also on the firm itself that these signals exert their influence. The firm had some role in determining the details of its output combination even within the framework of traditional economic management. And, following the reforms, its sphere of authority has grown further in this respect.

While emphasizing that the feedback control mechanism described above does *exist*, I must add that its effect is not very strong and only asserts itself with lags.[4]

[4]This was meant, among other things, in section 2.5 in which it was explained that demand, although an ineffective constraint in the instantaneous control of production, does affect production after all.

3.5. A few common problems of planning input and output combinations

After talking separately about the determination of input and output combinations, we now examine briefly one or two problems common to those two processes.

I have no intention of blurring over the important differences between the more and the less centralized forms of management, and between lower and higher degrees of enterprise independence. In the traditional economic management system the firm was given detailed instruction regarding the composition of its output. In numerous fields, its manufacturing technology was precisely defined. In addition, most of the inputs were centrally allocated, indicating not only what the firm was allowed to purchase but also from whom. Still, even in that situation the firm had a certain freedom (as regards the "finest" breakdown) in determining input–output combinations. Moreover, the views of the firm could also influence the formation of the central plan directive. Autonomy of the firm grew considerably in Hungary as a result of the 1968 reform, although high- and medium-level management authorities still exert a strong influence on the selection of input–output combinations even without plan directives.

Yet at this point what I wish to stress is not what renders the two economic management systems different, but their common features from the point of view of the subject of this chapter. In both situations high- and medium-level planning authorities as well as the firms themselves react to similar, *nonprice* signals.[5] In section 2.10 such indicators were called *shortage indicators*, z, and *slack indicators*, q. It is true, of course, that a comprehensive, regular and official observation of indicators z and q – repeated so as to be comparable – has not been organized anywhere. Yet a lot of written and verbal information flows between superior authorities and firms or between firms which may be considered in fact as a kind of approximation to the indicators z and q: they transmit information about the increasing or decreasing intensity of shortage and about the formation of slack (and within that, of mobilizable slack). *This highly complicated, in part officially recorded and in part informal, flow of information is fed into a multilevel allocation mechanism. This allocation mechanism ultimately directs resources away from places where shortage intensity is below its normal level and shifts them to where shortage intensity is above the normal level or where it is perhaps intolerably high.*

[5]The question of what extent the effect of price signals is added to these will be treated in Part II of the book. At this point we only state that nonprice signals play a very important role not only in the traditional but also in the postreform economic management system.

While emphasizing this common feature in the short-term planning of input–output combinations, the considerable difference existing between them must also be called to attention. In a resource-constrained economy feedback control affecting short-term adjustment of the *output* combination is much weaker than that which influences the short-term adjustment of the *input* combination. This follows logically from the mere existence of chronic shortages. It is very much in the proper interest of the producer to solve *his own problems*: if the supply of one input is permanently less adequate than that of some other input, the producer should somehow make a lasting adjustment to this relative scarcity. It is, however, *not his problem* but that of those using his products as inputs if the composition of his output is not appropriate. It is true that he is not indifferent toward complaints from the buyer–user firm and sooner or later will react to them in some way. But this reaction may be much slower and less reliable than the adjustment of inputs.

Because of all this – in spite of similarities and analogies – the two processes of short-term adjustment, adjusting the input and the output combination, are not equivalent. *In determination of the input combination the firm is comparatively more willing to adjust and be more flexible, it "gives in" more to circumstances; in determination of the output combination it is more rigid and more conservative, and it "dictates" more to the buyer dependent upon him.*

It is a recurring theme in this book that the *positions of the producer and user and of the seller and buyer are not symmetrical.* We shall talk about this in discussing trade ("sellers' market", "buyers' market"), in comparing the development of output and input stocks, in analyzing responsiveness to output and input prices, and so on. In any case, we have now made the acquaintance of one important manifestation of the asymmetry: the input combination reacts to nonprice shortage signals more sensitively than does the output combination.

3.6. On the motivation of firm managers

The behavior of the firm has now been described in two chapters. What we think about a firm's managers' motivations is now becoming clear to the reader.

There are many kinds of motivation behind human action. I would not like to trace them back – for an appearance of generality – to one single motive. It facilitates understanding of the behavior if we see what motives are simultaneously present and what conflicts may arise among them.

Not the only motive but among the most important ones in a firm's managers' behavior is their *identification with their own job*. I would not like to claim that the manager endeavors to do his *utmost* in the interest of the firm. Not all of them do, and then not on every day of the year and every hour of the day. Yet I would also refrain from the opposite statement. In standard microeconomics, the disutility of work is often mentioned, opposed to the utility enjoyable in consumption (and, of course, during leisure time). I think that this description does not apply to most kinds of work. And it is particularly unsuitable for describing the situation of those working in a managerial position. Direction, and decision-making in complicated situations are exciting tasks, and those doing it find a lot of problems, but also pleasure and satisfaction in it.[6] In the final analysis it can be said that on average a firm's manager tries *to do his job properly*, simply because a large proportion of people do so in most situations without any special motives.

Let us examine a little more closely how identification with the job appears. First of all in that the manager endeavors to secure *subsistence, survival, and viability* of the unit put in his charge. This implies that the output of the unit must be *acceptable* to the outside world. The hospital director tries to ensure that the practice of medicine should go well at the hospital; the school director ensures that teaching proceeds in an orderly way – that the patient and his relatives, or the pupil and his relatives should not be dissatisfied with the situation. The same "number one natural instinct" works with production managers: the subsistence, survival, and viability of the unit put in their charge must be secured, among other things, by making the output acceptable.

Another "natural instinct" is that the manager would like to guarantee for himself, and for the people under his direction, a *smooth* working process. He wishes to avoid confusion and disorder. If only for that reason he strives for the largest possible security: procurement of more input and larger reserves. (In subsequent chapters a third "natural instinct" will be treated: the expansion drive.) The effects of social conditions, material and moral incentives, ambitions and fears, rewards and penalties are then concentrated on these "natural instincts". With the exception of the owner-managed capitalist, the manager of every modern large firm depends on whoever appoints him. In the socialist economy now under examination superior authorities appoint the higher managers of the firm, and these in turn appoint the managers of lower ranks. Thus, a vertical chain of dependence is created. Under such circumstances it is a self-evident motive of the manager that he would like to win his superiors' acknowledgement, avoid their anger, and to fulfil their expectations: not

[6] See Zs. Hegedüs – Tardos (1974).

only their instructions but also their wishes. Central directives, expectations, and incentives may strengthen such action which would – even though perhaps more feebly – take place in any case. A bonus attached to the quantity of production can even strengthen the firm's endeavor to increase production. The incentives may also induce the manager to deviate from the elementary requirements of "work done properly". (For example, he may allow the quality of the product to deteriorate in seeking bonuses based on a bigger volume.) Such contradictions are amply discussed in the literature[7] and we need not go into details here.

Many people tend to believe that it is exactly here, in the bureaucratic dependence, that the motivations behind the quantity drive, rigidity in adjustment and, together with them, of shortage, must be sought. According to this all these phenomena appear because central economic policy enforces growth, wishes to submit everything to it and, by means of plan directives, drives firms to do the same. In my opinion it is *not* here that the *main* explanation for the group of phenomena under examination in this book may be found, although these factors may add to other effects. The motives enumerated above are *special* and assert themselves only in certain historical circumstances. Personally I lay stress upon the motive "identification with the job" because that is *general* and is present in all circumstances. And the remarkable fact is that this general motivation is *sufficient in itself* to bring about the almost insatiable demand of the firm for inputs and, as we shall see later, an unquenchable expansion drive. Thus, we have arrived at a stronger assertion than when we explained these phenomena from the special motivational effects of the traditional socialist economic management system.

As a matter of fact, the key question is not which motive *generates* quantity drive, insatiable demand of the firm, and expansion drive. The main problem is whether forces exist to act in the *opposite* direction, that is which would lead production managers to voluntarily *restrain* their demand for input and their expansion drive. The further developments in the book try to approach the answer to *this* really important question, step by step.

[7]See Kornai (1957, 1959) and from the more recent literature Bauer (1975a, 1978) and Laki (1978a).

CHAPTER 4

The buyer: Shopping process

4.1. Introduction

After examination of the firm as a *producer* in Chapters 2 and 3 it will be examined as a *buyer* in Chapters 4 and 5.

In standard microeconomics shopping is considered as the action of a single moment. The buyer – aware of his well-defined demand – meets the seller and the transaction comes about immediately. We consider shopping as a *dynamic process*, taking time. To make this idea tangible let us accompany the buyer on his shopping route. The problem is analyzed *at the submicro-level*: we shall deal with a single buyer and with only one of his purchasing actions. Continuing the example introduced in previous chapters we shall observe the *purchasing agent* of the screw factory who wants to buy steel.

Events are presented in figs. 4.1, 4.2, and 4.3. A shopping *algorithm* is formulated there, and is shown in the form of a flowchart, in a similar way to the description of computer algorithms. As a matter of fact, there is not such a complicated decision, which has to be made in an instant by a single "great choice". Almost all decisions are formed in the course of a multistep decision *process* which can be modelled in the form of a *decision algorithm*. The shopping algorithm that is described in Chapter 4 is a special case of the wider category of decision algorithms. On the diagrams diamond shapes represent "branch points": alternative situations or alternative decision possibilities face the buyer at these points. Circles indicate terminals: if the buyer arrived there, the process would end. Finally, rectangles symbolize points that are neither branches nor terminals. For easier reference the various parts of the diagram are denoted by capital letters.

Shopping is a complex activity for the firm; we shall be obliged to narrow the description in several respects, as follows.

(1) We shall deal exclusively with the purchase of materials, semi-finished goods, and parts (in what follows *material* for short) required for current production. Allocation of labor and investment goods will be analyzed in later chapters.

(2) Only storable *products* will be treated. The purchase of nonstorable *services* raises similar problems in many respects, but it has special features for the examination of which we have no space.

(3) The algorithm expounded below is a *general* scheme which could in fact describe the purchasing process of any buyer in any economic system. Yet in the interpretation of each step of the algorithm and in the formulation of illustrative examples we always had in mind purchases by a firm *in a resource-constrained economy*. There *suction* prevails in the trade as a means of production.[1] The buyer experiencing shortage tries to "suck", that is "siphon" to himself the desired input; the seller need not make much effort to press his goods on the buyer.

Because our discussion is general it is not important to specify at the outset whether we are dealing with the traditional or the postreform type of socialist economic management system.

Trade as the means of production takes place in a different way in the traditional socialist economic management system than it does in the postreform system.

For the traditional socialist firm a considerable proportion of materials is centrally allotted. The concrete system of rationing may differ by countries and by periods. Yet even with the highest degree of centralization official rationing determines only relatively aggregate quotas.

Rationing is anyway completed by the "business" contract between seller (steel-works or iron- and metalware trading firm) and buyer (screw factory), in which they agree on the specific quality, price, term of delivery, etc. Interfirm agreements are made all the more important by the fact that the system of central rationing usually covers only some materials, while the rest gets to its destination exclusively by way of direct contacts between selling and buying firms.

The Hungarian reform has almost eliminated central material rationing. Hence the role of interfirm agreements has increased. In any case, we do not deal with vertical central rationing at all in this chapter; horizontal interfirm relations will be treated exclusively. The question will be referred to again in the next chapter.

[1]The concepts of "suction" or "pressure" prevailing on the market were introduced in my book *Anti-Equilibrium* (Kornai, 1971a, b). In the present book I revert at several points to the description and explanation of the suction phenomenon.

From these introductory remarks, the shopping algorithm to be presented here can be interpreted in two different ways. *Either* our buyer has a valid rationing certificate in his pocket which, however, does not specify the source of purchase with obligatory force nor prescribes in full detail what the buyer should purchase; *or* our buyer is going to purchase a product which is not within the scope of central material rationing.

It is exactly these two interpretations that allow our "shopping algorithm" to relate equally to the traditional and postreform socialist economic management systems.

(4) For the time being we will not analyze how buying intentions, that is the initial demands introducing the shopping process, develop. That will be the subject of Chapter 5. In this chapter the initial demand is assumed as *given*.

(5) The role of prices will be treated in Part II of the book; here we will not touch upon it. Shopping processes take time and we assume that prices valid at the beginning of the process remain unchanged till its end. It is therefore not to price changes that the buyer's decisions react during his shopping route. In other words, we again have to deal with a "quantity" nonprice adjustment process.

(6) It is assumed that the quantity and composition of products available for buyers at the points of sale, that is *supply in this physical sense*, are also given. Actions taken by the buyer in question during the shopping process do not affect the physical volume and composition of goods instantaneously available from sellers. It is another question whether supply may finally react to this process (and to other buyers' similar shopping processes taking place simultaneously, and to the continual repetition of all these shopping processes), yet for now we shall leave this out of consideration.

(7) In section 2.2, talking about production, we discussed three degrees of adjustment. The same categorization will be applied now in the analysis of purchase. Sections 4.2 – 4.5 consider the following: *the buyer's instantaneous adjustment* to the supply which he finds in the course of a single shopping route. The buyer's short-term adjustment will be treated in the second part of this chapter, as well as in Chapter 5.

4.2. Beginning of the shopping process: Success or forced substitution

Let us assume that our buyer's initial demand for today is 100 tons of steel of quality 10. And now he sets out on his way. The point of departure is in *Field A* in the upper left-hand corner of fig. 4.1. He goes to selling point 1.

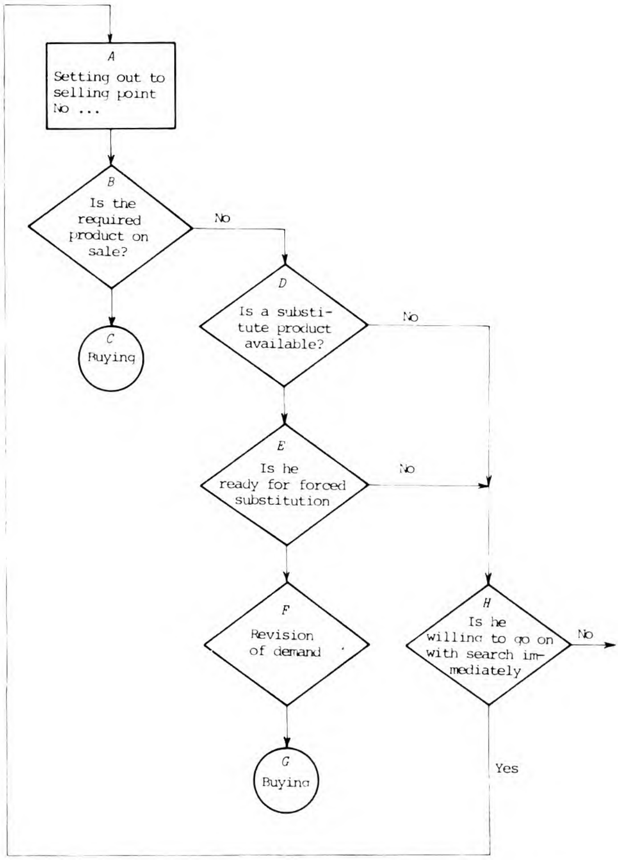


Figure 4.1. Shopping process: success, forced substitution, search.

(That is to the nearest-lying steel-works serving producer firms directly.) Is steel of quality 10 available today? (*Field B.*) If so, purchase can immediately take place successfully; initial demand has been satisfied. We have reached *Field C*, a terminal.

It happens, however, and quite frequently in a suction economy, that the material in demand is not to be had just then: *the buyer hits a supply constraint*. This is a phenomenon analogous to what was described in Chapter 2 in the following terms: *the producer hits a resource constraint*. In Chapter 2 it was in the *producer's* role, here in its *buyer's* role, that the firm is unable to carry out its initial intention (in the former case its intention to use an input for production, in the latter case its intention to purchase an input). There it was *within* the producer firm that the required input was not available, and here it is *outside* the firm, at the sales point of a seller.[2]

What happens if the buyer could not satisfy his initial demand at the first selling place? Let us assume that steel of quality 10 is not on sale instantaneously, but steel of qualities 9 or 11 is. If the purchasing agent buys one of these, he effects *forced substitutions*. (*Fields D→E→F→G.*) This is certainly forced substitution on the part of the *buyer*, since he deviated from his initial demand. It is another question whether this will also lead to forced substitution in *production*. If quality 10 is almost out of stock in the firm, then presumably it will. If it is still in stock, and within a short time a fresh supply is received, forced deviation from the technology specified in the plan (requiring steel of quality 10) may not be necessary. Let us remember what was said earlier: forced substitution in production either means that unnecessarily expensive, that is "too good" material is used, or that a cheaper and poorer quality is used than that prescribed, which damages the quality of the product. Forced substitution always causes loss – one way or another – to the buyer firm, who makes a concession to the seller by accepting the substitute product.

The distinction between voluntary and forced substitution which was already touched upon in Chapter 2 must be treated briefly once again. We can best understand this distinction in the present context if it is not a single purchase that we observe but a series of recurrent, regularly repeated purchases. Let us assume that at the selling point steel of quality 10 as well as quality 11 was on sale on every occasion. If the buyer had always been taking quality 10 and now, although both are available, he switched to quality 11, it would be a *voluntary* substitution. It is irrelevant what his motives for the switch might be: relative prices might have shifted; he could have been persuaded that the use of quality 11 would be better, and so on. The important thing is that *substitution was voluntary, since the other alternative was also physically available*. If, however, the changeover from 10 to 11 was caused by the fact that steel of quality 10 was not available at all, it is a *forced* substitution caused by the *physical absence of the other alternative*. This distinction is of great importance, not only within the narrow sphere of production and trade phenomena analyzed here. The two

[2]Although "hitting a supply constraint" and "hitting a resource constraint" are parallel phenomena, we have discussed the whole of the system in terms of the latter. The roots of the reproduction of shortage are to be found in the sphere of production and not in that of trade. It is physical resources available for production that represent the ultimate effective constraints of the system.

kinds of basic situation (voluntary and forced adjustment, and their various combinations) are met with in many spheres of society and fields of economic decision.

The actual performance of forced substitution is preceded – as is seen in fig. 4.1 – by a *revision of the initial demand.* The buyer may have set out already with a *demand strategy* determining several steps in advance: "If I get material of quality 10 I shall take it. If not, I shall try to get quality 9. If that is not available either, I shall try quality 11." Or, he may not have had such a definite strategy, intending to improvise after each failure, choosing from the available substitute products. In the next chapter this will be treated in more detail. So much, however, is clear: the *formation of demand is in itself a dynamic process.* The same buyer – in the course of the same shopping route – may have thought one thing at 9 o'clock in the morning but revised his opinion an hour later. And if at the same calendar time, for example at 9 a.m. on 17 May we observed the *degree of maturity* of demand, on the part of buyers interested in an identical group of products, we would find that it can differ widely between buyers. One group of buyers is making its initial demand, the second group has already revised it, the next group has done so twice, and so on.

Let us return to our algorithm. We assume that the buyer decides – weighing also the consequences – for forced substitution. Again we have reached a terminal: although deviating from the initial demand, a purchase has taken place.

The sequence of events will take a different turn if the seller has no substitute product to offer to the buyer, or if there is a substitute offered the buyer is not willing to take it. Before going into a discussion of the common continuation of these two possibilities let us examine them separately.

In the first case there was no substitute at selling point No. 1. (The process goes to the right from *Field D*: in the direction of "No".) The meaning of this statement depends on what we call a "substitute". Let us assume that the selling division of the steel works had no steel at all in the quality range 8–12. But it had reinforcing steel. This is also needed by the screw factory since it wants to do building work: it intends to enlarge its office building. In standard microeconomics reinforcing steel would in that case be considered a "substitute" for steel quality 10, since both are useful for the firm. In accordance with this, two products are mutually substituting if both have positive marginal utility.[3] My book does not accept this

[3]For criticism of the hypothesis of "general substitutability" see Hoch (1962). Similar thoughts are expressed in Lancaster's consumption model (Lancaster, 1957, 1966).

train of thought. The engineering factories waiting impatiently for screws, and not receiving them because the screw factory could not procure steel, will not be consoled by the news that material could be bought for the construction of an office building for the screw factory. What is more, the screw factory will not be satisfied either, since because of material shortage it cannot fulfill its screw production plan.

This is not terminological hair-splitting; the question is not about what we *call* substitution. The real question is whether one kind of advantage, success, or result *really* compensates for another kind of disadvantage, failure, or loss: "Your family suffered an accident – but you won the first prize in lottery...", "Flood has swept away riverside villages – but a new factory was built". The striking absurdity of these "compensations" indicates that it is not possible to add up "pluses" and then deduct from them the "minuses". The general spirit of either an individual or the nation or of any kind of group is no pocket calculator that finally counts the net balance of the positive and negative sides. The individual, and also the community, experiences the sensation of different events separately: the good and bad, success and failure. The question leads us to the foundations of utility and social welfare theory. Later on we will deal with these repeatedly. We only intended now to make it understood that we have a good reason for insisting on a narrow interpretation of substitution.

For the time being we need a pragmatic delimitation: how far should the concept of "substitution" extend; what is to be considered substitution and what goes beyond it.

Two inputs (or two kinds of input combination) are mutually substituting for each other if either can be used for producing identical (or approximately identical) outputs. That is, for example, a certain range of screws can be produced from steel of quality 9 or 10. The adjective "identical" with regard to output was not qualified without reason. It is still substitution if the output produced by one input combination is not perfectly identical with another: there is, for example, a slight difference in quality – as long as the buyer can be forced to accept goods of reduced quality. This delimitation is relative, to a certain extent; it depends on the relative powers of seller and buyer. As long as – even though reluctantly – the engineering factory accepts screws made from the poorer steel of quality 9, then this quality is a substitute for quality 10. If, however, the engineering factory rejects the screws material of quality 9 does not substitute for quality 10.

Although the logic of this delimitation as given above is clear, there may be doubtful border-line cases in practice. In such cases the following recipe should be used. Let us forget what the jargon of standard microeconomics has accustomed us to in the interpretation of the word "substitution". If it is a production input, let us pose the question: would the engineer consider B a substitute for A? If it is a consumer good, would the housewife call B a substitute for A? And, if the engineer does not consider reinforcing iron a substitute for steel of quality 10 and the housewife does not consider the theatre ticket to be a substitute for meat, let us not consider it so, either.

In the second case although the seller has available a supply of the substitute (e.g. steel of qualities 9 or 11 instead of 10), the buyer is not willing to undertake forced substitution. (The process goes to the right from *Field E*, in the direction of "No".) This proves that he is not acting *entirely* by force of circumstance, but still has a few alternatives.

4.3. A continuation of the shopping process: Search

Now let us return to the buyer who is at the arrow pointing to the lower diamond on the right-hand side of fig. 4.1. No forced substitution has taken place since there was no possibility or willingness. The next dilemma is whether the buyer should go to another source of supply to see if the product he initially demanded is available there. Let us assume, for example, that seller no. 1 is not in a monopolistic position: other sources of supply exist. (For example, another steel-works situated farther away, or the various depots of iron- and metal ware trading firms, etc.)

The buyer has decided to begin *search*. (He proceeds downwards from *Field H*, in the direction of "Yes".) In the language of computer algorithms he starts a new cycle: he returns to the uppermost rectangle, which now receives a new serial number. The buyer sets out to selling place 2, where everything begins again. And it may happen that further cycles will follow: the buyer goes to selling places 3, 4, and so on.

At some point he will stop. Either because he finally obtains what he wanted, or because he grew tired of searching and finally brought himself to accept forced substitution.

Search activity involves various sacrifices; time, physical effort, or perhaps other inputs (transport, post, telephone, etc.) are required by search, not to mention nervousness and vexation.[4]

[4]The description of the shopping process touches at this point the problem-setting of the so-called "search models". These describe various kinds of search processes, for example the worker searches for employment. (See Phelps, 1970b; Holt, 1970; Lippman–McCall, 1976.) The buyer searches for the cheapest among products offered at many places at different prices, etc. Concerning the latter topics see the survey by Rothschild (1973). In Sweden models have been elaborated for the analysis of the search for flats. (See Gustafsson–Hårsman–Snickars, 1977.)

Let us now take the case in which the buyer first tried to be stubborn: he insisted on his initial demand at the first and perhaps even at the second attempt to buy. But failure may soften his resolution. To selling point 3 he already sets out with the thought: "If possible, I shall buy quality 10, but if it is not available there either, I shall take quality 9." That is, in this cycle he does not set out with the original initial demand, but with a *revised* one. The same is seen here as was mentioned in connection with forced substitution: in the dynamic interpretation of the shopping process there is no one specific "demand". *Demand formation itself in the framework of instantaneous adjustment to supply is a dynamic process.* As a result of experience gathered in the course of the process it may be modified from its initial state.

4.4. Alternative continuation of the shopping process: Waiting

The upper left-hand corner of fig. 4.2 shows *Field H* as seen from the lower right-hand corner of fig 4.1. Is the buyer willing to go on with his search without delay? If so, he will return to the initial point of the cycle presented in fig. 4.1, that is to *Field A*. If, however, he is not prepared to go on with his search immediately, what can he do? *Wait.*

Waiting takes time just as search does. The difference is that search is an *active* reaction on the part of the buyer, while waiting is a *passive* one.

He waits for a certain time, for example, a week. (See *Field I*.) If that period of tolerance has passed, he can reconsider his opinion (*Field J*): should he wait another week? If so, he is caught up in the cycle once more: he steps back to *Field I*. If, however, he does not want to wait any longer without doing anything, he brings himself to make inquiries (*Field K*), whether the required material has arrived at the selling point. In other words, he starts again on the cycle shown in fig. 4.1, from *Field A*.

This *repeated inquiry* is a phenomenon resembling search. "Search", in the preceding description, meant that the buyer visits *various* selling points in turn. This he can do in a short time, or he can even make a round of telephone calls at one time. As opposed to this, repeated inquiry with the *same* seller corresponds to "search" spread over time.

Search divided between different selling points and repeated inquiry at the same point have a common feature: the buyer gains *information*. Search and repeated inquiry could be reduced or even cease if the buyer were given the information concerning where and when the required product is to be had. *In circumstances of shortage a considerable part of this information-gathering activity falls on the buyer.*

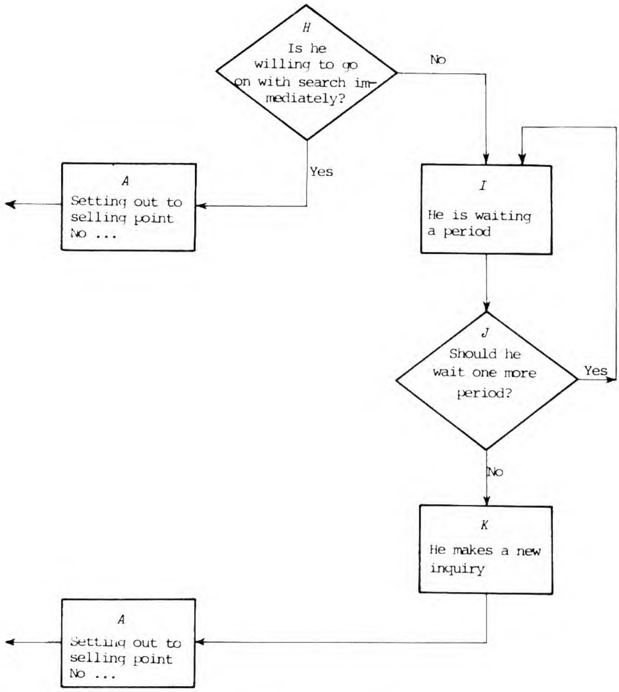


Figure 4.2. Waiting.

4.5. Queuing

A special case of waiting is *queuing*. It is possible that already at the first selling point, where the buyer called first, one had to queue for the product. Therefore, our buyer who submits his claim to the seller on Monday morning will be served when the demanded product arrives, ahead of other buyers who announced their purchasing intention on Monday afternoon or Tuesday. Occasionally queuing is real (in material purchasing this is rare). In other cases the queue is only imaginary: it is the assertion of the "first come first served" principle, that is the one who makes his claim earlier will get his share of the material instantaneously out of stock sooner, when it arrives.[5] Waiting *may* be combined with

[5]Queuing has an extensive mathematical literature. See, for example, Cox–Smith (1961), Jaiswal (1968), and Knudsen (1972).

queuing, yet a queue is not necessarily formed in every case. A material of uncertain supply – first in stock, then out of stock – can be sold in such a way that when it reaches the seller he will give it to the buyer who shows up first. If the buyer is lucky (or he inquires frequently), he will receive the material; if he is unlucky or only visits the seller rarely, another buyer will take the goods from under his nose.

Waiting combined with queuing cannot be drawn as a branch of *inactive* waiting in fig. 4.2. In fact, fig. 4.1 representing the original shopping cycle needs to be extended by a few additional blocks. This extension is provided by fig. 4.3. The figure speaks for itself, and we only draw attention to *Field M*. How many other buyers precede the buyer in the

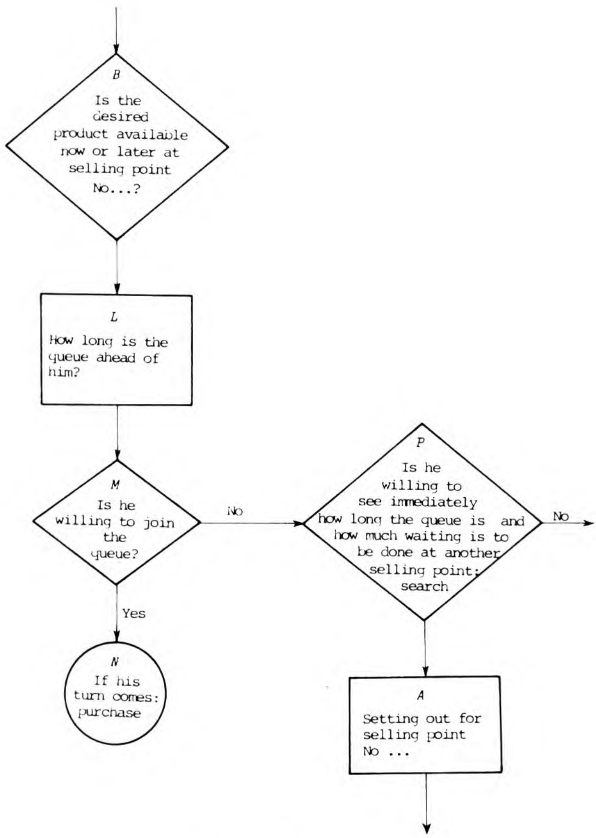


Figure 4.3. Queuing.

queue, how long a queuing time can he expect? These questions may strongly influence the buyer in his decision whether to join the queue or not. We return to this problem in Chapters 7 and 17.

4.6. Efforts to win over the seller

The shopping process has been surveyed up to now in chronological order of decision-making. In the course of the discussion it was tacitly assumed that the question was exclusively whether the demanded product was available at a certain place and time. If it is available, the buyer secures it, if not, then he does not receive it. Yet real life often produces another situation. The product arrives at the selling point. In a suction economy, under conditions of chronic shortage, the seller knows that several buyers aspire to it and he has to decide to which of them he should give it. Whom should he advise that the goods have arrived? In practice material purchase does not usually take place in such a way that the buyer personally visits the seller and has the goods brought out for himself. Actual transaction is often preceded by correspondence, telephone calls, and repeated personal discussions. The seller knows if there are many claimants to the product, so that he can select from them. *The buyer also knows that the seller will select from among the buyers.*[6] *Therefore, he makes efforts to win over the seller*, so that he should be preferred to other buyers. There are several methods to achieve this aim.

(a) Many sellers give priority to a regular customer over others. Therefore the buyer tries to be loyal.[7]

(b) The seller may prefer a buyer with whom there is no problem, who is not making difficulties. Therefore it is worthwhile for the buyer to think twice before he makes a complaint to the superior authority or he asserts his rights according to the terms of the contract, for example he demands the payment of a penalty.[8] He may obtain justice in this particular case,

[6]On the "competition of buyers" see the pioneering study by Péter (1956), and the later works by Falus–Szikra (1974a, b, 1975) and Román (1973). In Chapter 17 we revert to the special "rationing" form in which the seller enjoying a monopolistic position distributes shortage goods among buyers.

[7]Again I refer to the book *Exit, Voice and Loyalty* by Hirschman (1970). In a shortage situation the buyer tries to be loyal to the seller and does not avail himself with the means of "exit" even in the case of complaints. Even if there is another seller, it is in many cases much more helpful to be tolerant and make use of the advantages due to the "loyal buyer".

[8]Continuing Hirschman's train of thought, he is deprived to a certain extent not only of the means of "exit" but also that of "voice".

but next time the seller may take his revenge, and in a way that would be impossible to detect, so that the buyer would not be in a situation to make a complaint again. (For example, the seller does not advise him about the arrival of a long-awaited product, or lets him know later than his competitor.) What is more, the passive attitude ("I do not make a complaint") is not enough. An active attitude is better: a definite effort to be "good friends" with the seller. It is useful to establish a personal friendship, since this also brings advantages with business contacts.

(c) While paragraph (b) mentioned how the buyer wins the seller's favor by friendly *words*, the other possibility must also be mentioned, namely that the seller can be won over by *deeds*. Simple corruption occurs, too: the buyer from the screw factory bribes the stock-keeper of the steel-works to telephone immediately on arrival of the long-awaited steel of quality 10. Or he pays something to the sales department man: if there are several claimants, he should be given the product and not someone else. Bribery sometimes involves money and sometimes takes the form of some "present". This is, of course, illegal. Such corruption is sometimes exposed, followed by the appropriate legal sanctions.

There are, however, numerous indirect forms of recompensing favors which are impossible or almost impossible to prosecute in law. The simplest is the case where seller and buyer are exchanging roles from time to time: today I help you out with steel, tomorrow you help me out with screws. In other cases it is not the one firm that compensates its partner, but the man in charge compensates the other man in charge. "Today you give me material; tomorrow my wife who is secretary at the district clinic will help you get to the doctor in advance of your proper turn." This exchange of services is not always that direct, but often has two or three intermediate steps. "My colleague's colleague will help your colleague's colleague...." *Every "buyer" is a "seller" somewhere.* Everybody who has received a favor somewhere is able to return it somewhere else. And, if the chain of mutual favors is well established, it may considerably influence the selection processes.[9] This chain of mutual services is much more important than the common direct corruption which is easy to detect and therefore too risky.[10]

[9]Using the ingenious analogy of Hankiss (1978): a "bank of mutual favors develops, in which every member of the community of interests places his deposit and receives in exchange an ideal 'letter of credit'...This letter of credit can be cashed not only with the person for whom the favor was done, but with every member of the community of interests to which both belong."

[10]K. Polányi (1944, 1976) holds *reciprocity* an allocation form of no smaller importance in human history than the *market* allocation form of "give and take for money".

(d) With a few services accompanying trading, it is not clear whether the seller or the buyer should take responsibility. For example, who is concerned with transporting the goods? In order to win the seller's benevolence the buyer may offer to take these charges on himself.

Virtually all the phenomena discussed here are observable. This may be difficult with some of them, but the observant researcher can describe them in case studies; for example, the manifestations of "mutual favors". Other symptoms could be quantified which would enable testing of the hypotheses enumerated above. For example, data can be compiled about the buyer's officially raised complaints, or about the sharing of transport costs between buyer and seller, and so on.

Later we will discuss the symmetrical phenomena, namely the efforts used by the seller to win over the buyer. But it is already clear at this point that in the act of trade it is not just *objects* exchanging ownership, but also a *contact between man and man is established.* And this is not only the personal relationship between the seller M.M. and the buyer N.N., but a *permanent social relation between two social roles.* This relation is very dependent on how the relative strengths of the two partners develop: whether there is a "seller's market" or a "buyer's market", and which one of the trading partners is in a situation to dictate and which has to obey.[11]

4.7. The buyer's attitude

In sections 4.2–4.5 we accompanied the buyer on a single shopping route. In most cases, however, the process *repeats itself* from time to time. It is, of course, not certain that the buyer – faced with a similar dilemma – will again decide in the same way, yet it is assumed that his decisions will show a certain stochastic regularity.[12]

Let us take, for example, the branching in the lower right-hand corner of fig. 4.1. The buyer did not find the material required, and he cannot carry out forced substitution or does not want to. Should he set out to find another selling point immediately? The buyer's *search propensity* is shown on fig. 4.4.

[11]On the relative strengths of the buyer and seller see Fábri (1973).

[12]In wording the text of section 4.7 I made use of ideas which developed in the course of joint research with Jörgen W. Weibull and of which Chapter 7 of the book and Mathematical Appendix A give a detailed account.

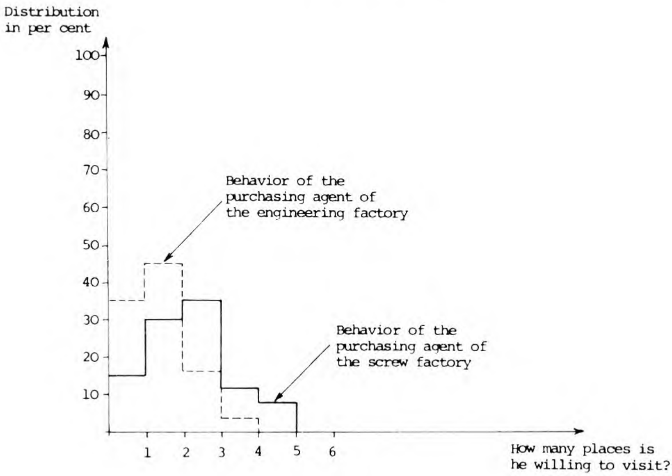


Figure 4.4. Search propensity.

Let us assume that the buyer is looking for a single item of some spare part. The horizontal axis shows the number of places the buyer is willing to visit. What is shown is not the number of places he actually visited. If, for example, he was lucky in five cases out of ten and received what he wanted right at the first place, nothing can be said in those five cases about how many places he would have been willing to visit in the search process. Only decisions related to failures to purchase need to be considered here. On one occasion, not being able to get the spare part initially demanded even at the third place, the buyer decided not to go on to the fourth place but to accept forced substitution instead. In this single case his search propensity was 3. On the vertical axis we can read how often, expressed as a percentage, his search propensity took particular values. According to the figure in 15 cases out of 100 the buyer, the purchasing agent of the screw factory – represented by the continuous line – stopped at the first place, in 30 cases he ended his search only at the second place, and so on.

Another buyer reacts differently to a similar dilemma. This buyer, the purchasing agent of an engineering factory – represented in the figure by a dotted line – is more impatient: in most cases he stops at the first or second place, and is never willing to try five places.

The two types of buyers may be characterized by the *intertemporal average* of their repeated reactions: the average search propensity of the first buyer is 2.68 selling places, while that of the second, the impatient one, is only 1.89. A fuller description is provided, however, by giving the statistical *distribution* of alternative reactions, as shown in fig. 4.4.

Search propensity is a component of the *buyer's attitude*. This attitude is the buyer's subjective response to recurring situations, and sums up various propensities, willingness, intentions, and aspiration. In the course of a single submicro-level series of events or along a single shopping route the buyer only puts into effect his instantaneous adjustment to the supply of that day. (First degree adjustment.) Buying attitude gives account of a *more permanent* adjustment, i.e. of the relatively more constant features of behavior. This is, however, still only the second, *short-term* degree of adjustment. The third, *long-term* degree is the gradual shift of attitude resulting from secular changes in the nature of the purchasing situation.[13]

We will enumerate the most important components of the buyer's attitude in more detail, but before doing so we make a few methodological comments on their nature.

Strictly speaking, we ought to use *random variables* for the description of all the components of attitude and characterize the attitude by their joint distribution. At this point, however, we do not wish to make use of this method of description, since our intention is only to illustrate a few general ideas. Therefore, we content ourselves with a rougher but simpler deterministic description in which only the expected values of random variables appear.

One component of the attitude can be expressed either by a single real number, or by a vector. For example, it is expedient to represent the search propensity by several indicators: how many places the buyer is willing to visit, how long he is willing to search, or perhaps how much money he is disposed to spend on search, and so on.

The buyer's attitude variables may be influenced by certain factors (apart from his own hesitation and inconsistencies). Let us look at an example to illustrate this. One attitude variable is the *queuing propensity*. (See *Field M* in fig. 4.3.) As already mentioned, this buyer's decision depends on the expected queuing time. In the case of such a relationship the following expression will be used: the attitude variable in question is a *function of* this or that *explanatory variable*.

In the final account, therefore, the *buyer's attitude can be described – in regard to its mathematical form – as an ensemble of parameters and functions*. This mode of description will be illustrated by a formalized example in Chapter 7 and also in Mathematical Appendix A.

[13]Short- and long-term adjustment of demand will be discussed in the next chapter.

We now summarize some of the main components of the buyer's attitude, arranging them according to their economic content.[14]

(1) *Initial demand function.* The explanatory variables on which this function depends will be discussed in the next chapter; here it is mentioned only to make the survey of attitude more complete.

(2) *Forced substitution propensity.* This can be described, for example, by specifying the fraction of his initial demand for which the buyer is willing to purchase a substitute. It may be useful to break it up into several variables, distinguishing between propensity to make "small" and "large", "near" and "distant" substitutions.

It may be a function of the search preliminaries: the buyer will probably be less stubborn at the fifth selling point than he was at the first. It may also be a function of the relative prices of the inputs initially demanded and the potential substitutes.

The forced substitution propensity is inseparable from the buyer's willingness to *revise* his initial demands. Perhaps the dynamic regularity of this revision could be described. For example, the buyer sets out with a "finely" broken down, disaggregated demand, and in case of failure he changes over to a "less fine" and more aggregated demand. (He sets out with "I am looking for steel of quality 10" and, after revision, "I am looking for steel of quality 9–11".)

(3) *Search propensity.* This was discussed in detail as an expository example at the beginning of the present section.

(4) *Queuing propensity.* This may be a function of the expected waiting time, among other things.

(5) *Efforts used to win over the seller.* This may be a function of the duration of search and of the length of the queue, and so on.

I do not claim that the few components mentioned above are sufficient to describe fully every aspect of the buyer's attitudes. The picture can obviously be further enriched, particularly if we draw into the analysis the buyer's reactions to purchasing prices.[15]

[14]With one exception all components appear in the diamonds of figs. 4.1–4.3, since these components express the buyer's attitude to one or other decision problem. The exception is the fifth component, the buyer's effort to win over the seller. This is not a separate "diamond" because it accompanies the whole shopping process.

[15]Two examples are mentioned. In Chapter 14 we will analyze the firm's responsiveness to input prices. In Chapter 15 the question will be treated, as to how far the firm as buyer resists possible efforts by the seller to increase prices. Indicators describing these two essential features of the behavior of the firm obviously belong to the description of the buyer's attitude.

CHAPTER 5

The buyer: Demand formation

5.1. Introduction

In the first part of Chapter 4 we accompanied a single buyer as he set out on his shopping round with a given shopping list in his hand. The main subject of the present chapter is how this shopping list developed. What are the permanent features of the buyer's behavior that determine initial demand in the continuous repetition of the shopping process?

Most of the assumptions described in section 4.1 – serving the delimitation and simplification of the discussion – remain valid. Using the same numbers as there: assumption 1 (we shall deal exclusively with the purchase of materials; employment of labor and acquisition of investment goods will not be covered); assumption 2 (only storable products will be treated); assumption 5 (effect of input prices will be disregarded); and assumption 6 (supply will be considered as given) are all valid.

For most of the chapter assumption 3 will also remain valid. Thus, we shall deal with the demand of the socialist firm without distinction, from the point of view of the present chapter, between the traditional and the postreform economic management system. Such a distinction will be made only in the last sections of the chapter. That is, until then we shall disregard the role of central material rationing and material quotas. The concept of material demand of the firm is interpretable in this way, too, and its formation can also be explained if we content ourselves with examining the *horizontal* relations between buyer firm and seller firm. The *vertical relations* between the superior authority rationing materials and the firm first claiming and then using such material will be treated in section 5.7.

5.2. A simple rule of thumb

We will use an example by way of introduction. It is deliberately presented in the simplest form possible.

Determination of the initial demand for material purchase is a permanently recurring and habitual routine task. The purchasing agent of the screw factory in our example composes his shopping list using the following rule of thumb.

Let us buy as much of each of the materials in question as will take the existing input stocks of the materials up to the *desired level.*

Which materials come into question? First of all those expected to be used within a reasonable time (e.g. within 12 months), assuming that the technology specified in the plan will be used and output produced in the planned composition. In addition, the firm will also be concerned with the purchase of materials which may facilitate instantaneous forced adjustment (forced substitution, modification of the output composition) within the time horizon.

The desired level of input stocks is generally three months' use of material. Exceptions are materials of uncertain supply: for them, the desired level is six months.

To apply the above-mentioned rule the purchasing agent must obtain information about the expected use of material. The man from the screw factory first tries to find out the following: what is going to be the average monthly use of steel of quality 10 or 11 during the coming months if the firm produces the planned volume by the technology specified in the plan, with the planned input–output combination. After that there follows the consideration of how much more steel of quality 11 would be used if steel of quality 10 were not to be had. But also the reverse case is considered, namely how much more steel of quality 10 would be used if quality 11 were to become a shortage good.

The final result of these considerations is the *expectation*[1] of future use of material. At this point it would serve no purpose to define this volume exactly, since the purchasing agent has not calculated it with the aid of some complicated stochastic inventory model. It is the result of very rough calculations in which the guiding thought is this: "let us be prepared for the worse case". I think that the meaning of the expression "expectation of future use of material" can also be perceived in this way. This is the conditional forecast of the purchasing agent made in conjunction with the production managers: approximately so much of the *i*th material will be

[1]Expectations are the subjective reflection of some future process in the decision-maker's mind. It is the reflection upon which the decisions and actions of the decision-maker are built.

used if instantaneous conditions direct material consumption towards the ith material.

However simple the rule of thumb may seem, its application means in practice the execution of a three-step decision algorithm.

First step. Draw up the list of materials eligible for use.

Second step. Determine the expected volume of use of eligible materials.

Third step. Calculate initial demand relying upon formula (5.1) below. The formula gives the ith component of the vector of initial demand, obviously assuming that the ith material is among those eligible for use.

$$\boxed{\begin{array}{l}\text{initial}\\ \text{demand}\\ \text{for the}\\ i\text{th}\\ \text{material}\\ \text{on day } t\end{array}} = \begin{cases} \boxed{\begin{array}{l}\text{coefficient}\\ \text{of the}\\ \text{desired}\\ \text{stock}\end{array}} \times \boxed{\begin{array}{l}\text{expected use}\\ \text{of the } i\text{th}\\ \text{material for}\\ \text{one period}\\ \text{of time}\end{array}} - \boxed{\begin{array}{l}\text{input stock}\\ \text{of the } i\text{th}\\ \text{material}\\ \text{on day } t\end{array}} & \\ \boxed{\quad\text{desired stock}\quad} - \boxed{\text{actual stock}} & \text{if actual stock is below the desired} \\ 0, & \text{if actual stock is not below the desired} \end{cases} \tag{5.1}$$

The rule of thumb, together with the algorithm describing its implementation, including the application of formula (5.1) as a third step, present together a concrete illustrative example of what we may call more generally the *material demand function of the firm.* Those accustomed to the neoclassical form of demand functions (demand as a function of price and income, and, what is more, as an exactly formulated and analytically convenient function) may find both the form and the content of the description surprising. A comparison of its *content* with the neoclassical demand function will follow later. At this point I wish only to justify the *terminology*.

Sticking to our example: the rule and the associated algorithm ensure that a definite relationship should obtain between some explanatory variables and the dependent variable (initial demand). Consequently, we are certainly dealing with a function-like relationship. "Function", after all, is a very wide category, and it is not reserved for the functions convenient in mathematical analysis. Since the relationship in question is one that is rather intricate in reality, I do not endeavor to make it seem too simple for the sake of analytical "beauty".

5.3. Material demand function of the firm

Now let us turn to a more general discussion. Which features of the rule of thumb and algorithm described in the preceding section are *incidental* and only of illustrative character and which prove to be *general* features of the material demand function of the socialist firm will gradually become apparent. It will facilitate understanding if we make some comparisons with the neoclassical demand function as the latter appears within the framework of a general equilibrium model.

(1) In general equilibrium theory the volume of production of the firm, as well as the actual input–output combination, are formed *at once* and *simultaneously*. The relative prices of both inputs and outputs are taken into account, the budget constraint of the firm is observed and, upon this basis, the input–output combination is determined which ensures achievement of maximum profit for the firm. Accordingly, the firm's material demand is unambiguously determined.

As opposed to this we interpret the formation and mutual adjustment of both producing and purchasing intentions as a *dynamic process*. The continuous and intricate mutual adjustment processes are divided, in our abstract theoretical framework, into two time phases. One phase is the determination of the *short-term* production plans of the firm.[2] This was discussed in Chapter 3; it was there that we discussed the factors influencing the planning of short-term aggregate production, technology, and the composition of output.[3] A particularly important role is played in this by shortage signals which influence plans concerning the composition of both output and input. Persistent shortage and excessively frequent interruptions in supply in practice prevent the buyer from using certain inputs; technology and the output combination will be adjusted in response to these chronic shortage phenomena.

The other phase is the *instantaneous* adjustment of the firm to currently existing conditions, in the role of both producer (Chapter 2) and buyer (Chapter 4). It is in the latter role that its instantaneous buying intention is formed. This is, therefore, a *secondary* phenomenon deriving from the primary phenomenon, i.e. the already given short-term production intention.

[2]In Chapters 4 and 5, dealing with trade as well as with demand and supply, we do not dwell upon questions of long-term control.

[3]For the time being – up to Chapter 14 – we put aside questions of the role played by relative prices in demand formation. They have a certain influence on the traditional socialist firm and their impact may be somewhat strengthened in the postreform economic management system. This influence is, however, weak enough for us to disregard it for the present discussion. Therefore, it is not the problem of price and nonprice signals that we shall stress in drawing a comparison with the neoclassical demand function, but other differences.

(2) The neoclassical model of a firm's demand assumes tacitly that the supply side will present no obstacle to the satisfaction of demand. The general equilibrium model, at a high level of abstraction, assumes instant and perfect adjustment of supply. This can be translated into the language of a practical interpretation, with a certain looseness, as the requirement that there is an output stock of every product which usually serves as an input. When any particular output stock is depleted, new supplies can be obtained quickly. If the system is in a state of Walrasian equilibrium, there is never any shortage of anything.

We cannot apply this assumption. On the contrary, we have to describe exactly how chronic shortage affects not only the actual purchase, but even the purchasing *intention*.

Formation of the initial demand is based upon expectations. The firm as buyer is aware that its expectations are uncertain. It is compelled to take into account that production may not be at the planned volume or with the planned input–output combination. It has to consider also that there will be interruptions in material supply. That is why it develops a certain *safety strategy*. This may take many forms and in the preceding section we presented only one specific example of the many different existing and observable safety strategies. Purchasing rules of thumb are different between firms; algorithms serving their implementation are different; and the constant parameters used for calculating the actual shopping lists are also different. The length of time horizon considered in material purchase also varies. At numerous places the different materials are put into a number of "uncertainty classes". Fixed coefficients for the calculation of the desired size of stock are not used everywhere. But even if a simple linear form similar to (5.1) were applied, the expectations of the future use of material could be determined in many different ways. As regards the result of the calculation, that is the initial demand, in some places this does not go down to the "finest" disaggregation, but leaves more scope for improvisation on the spot. (Obtain steel of quality 9 or 10 or 11.) Or, from the beginning it might consist of a multistep instruction. ("Try to get quality 10; if you cannot, buy quality 11, etc.")

Possible variations could be enumerated much further. But, even though actual purchasing strategies differ from each other, it is their general characteristic that they wish to satisfy not just the instantaneous material demands of production. They seek to ensure some security of material supply by accumulating adequate input stocks. It is also their general feature that the greater the uncertainty, the larger the input stocks they try to accumulate. This interrelation is summed up in formula (5.2) below, which is related to a single kind of material: the ith material.

expectations concerning the purchase of the *i*th input:

- the longer the expected time between two dates of delivery of new supply;
- the greater the variability of the time interval between deliveries;
- the more frequently shortage occurs at the moment of purchase;
- ⋮

expectations concerning other elements of production and purchase:

- the more uncertain whether planned input–output combination will be observed in production;
- the less the *i*th material is substitutable;
- the larger the probability of shortage of materials appropriate for substitution of the *i*th material;
- ⋮

→ the larger the desired input stock of the *i*th material (5.2)

The three dots in the preceding scheme indicate that our listing of the "uncertainty factors" is not complete; moreover, uncertainty that has the effect of increasing the desired stock may also appear in other interrelations. In the literature of operations research there are familiar inventory models which determine the desired size of input stocks for a firm in relation to one or the other aspect of uncertainty.[4] At this point we do not attempt to give an exact description of these interrelations but rest content with the qualitative presentation of the direction of dependence.

(3) Scheme (5.2) leads us to the "stock-flow" problem. The neoclassical demand function describes a *direct* causal relationship between certain explanatory variables and material demand as a flow variable:

explanatory variables → material demand (*flow* variable) (5.3)

[4]See, for example, Prékopa (1965b).

We describe, however, an indirect causal relationship; *stock* variables are interposed thereby giving the desired size of input stocks:

$$\boxed{\text{explanatory variables}} \xrightarrow[\text{rule}]{\text{first}} \boxed{\text{desired input stocks } (\textit{stock} \text{ variables})} \xrightarrow[\text{rule}]{\text{second}} \boxed{\text{initial material demand } (\textit{flow} \text{ variables})} \tag{5.4}$$

In our view the firm's "material demand function" is composed from two rules. The first rule describes a relatively permanent feature in the firm's behavior, that is one component of its attitude as a buyer: what input stocks does the firm try to build up. And the second rule formulates the following: if the desired level of the input stock is already given, what will the instantaneous buying intention of the firm be?

(4) Finally, a further important difference is that here we are not talking about "demand" in general, but expressly about *initial* demand. If it is not satisfied, the buyer makes a number of further steps: he applies forced substitution, he searches, waits, and so on.

By now it should be clear why the demand function in Chapter 4 was called *one* of the components of the buyer's attitude, emphasizing at the same time that a number of other important components exist. We used the term "attitude" to mean all that was permanent and stable – at least in the stochastic sense – in the buyer's behavior. His "safety strategy" for purchases is constant with its main qualitative features and a few parameters. Yet these permanent behavioral patterns describe the whole of the buyer's attitude only *in association with other components*; we also have to specify how the buyer reacts to perceptions of shortage, what his propensity to accept forced substitution might be, and so on.

In the analysis of an economy in which shortage is a rare event, it is not so important to be concerned about these other components of the buyer's attitude. In other words, it does not matter much if the buyer is described solely by his demand function. After all, the important thing there is that if he is prepared to pay the price the seller asks, purchase will take place. If, however, we stopped at this point – describing only the demand function – also in a shortage economy, we would have said very little about the buyer's attitude.

I must add a remark of self-criticism, relating to the comparison with the neoclassical demand function. I have frequently used the words "demand" and "supply" in earlier chapters and will go on using them. In this regard I abandon the terminology of *Anti-Equilibrium* in which I tried to replace these terms by new ones; here I have returned to the traditional economic usage. Today I believe that here, as well as with a few other concepts, *Anti-Equilibrium* tried in vain to assert a kind of "terminological radicalism". It is not in the choice of words but in what we have to say that we must radically depart from what is old, as soon as this becomes necessary.

I would still maintain what I said in my earlier book about the lack of theoretical clarity in the neoclassical interpretation of "demand" and "supply". The static approach of standard microeconomics, and its neglect of the study of persistent deviations from the Walrasian equilibrium state leave numerous aspects of purchasing and selling processes extremely unclear and within these the dynamics of demand and supply. We must strive to present a more exact description of all this, but we need not discard the words "demand" and "supply". It is true that each designates a comprehensive group of phenomena, to which belong a series of specific phenomena. Specification is, however, better done not with the introduction of a brand new term, but with the addition of complementary attributes to the old basic concept. This is done in the present book when we discuss, for example, *initial* and *revised* demand.

5.4. The Clower interpretation of demand

It would not be correct to limit the argument to the traditional versions of neoclassical microeconomics. In the past fifteen years a new trend has developed: several economists began to deal with phenomena that appear when the market is out of Walrasian equilibrium. Pioneers of this important course are Clower and Leijonhufvud.[5] Outstanding stages in it are marked by the works of Barro and Grossman.[6] The theoretical apparatus based on this trend has also been applied by several authors in the analysis of a socialist economy.[7] In what follows I call this line of research the *Clower–Barro–Grossman school*.[8] The message my book seeks to convey is

[5]See Clower (1965) and Leijonhufvud (1968). I must make a personal remark here. At the time of writing *Anti-Equilibrium* (Kornai, 1971a, b) I was not familiar either with Clower's article, or Leijonhufvud's book. I developed my ideas independently of them, including the idea that in a state away from Walrasian equilibrium it is *either* the buyer's buying intention, *or* the seller's selling intention that becomes the effective constraint. (See Kornai, 1971b, pp. 252–256.) Generalizing the experiences of a socialist economy, my thoughts took the same direction as these of Clower and Leijonhufvud.

[6]See Barro–Grossman (1971, 1974, 1976).

[7]See, first of all, the various valuable studies of Portes–Winter (1977a, 1977b, 1978).

[8]A series of important studies can be listed here, among them the works of Drèze (1975), Grandmont (1977), Benassy (1974, 1975, 1977), Malinvaud (1977), Svensson (1977), and papers in the volume edited by Schwödiauer (1978). In several works of this school their findings are called "disequilibrium theory". For reasons to be explained later, I would not like to adopt this terminology. I think that the names of Clower, Barro and Grossman clearly mark the school. Every work in this list can be simply recognized by the fact that in their surveys of the literature, reference is always made to these pioneer authors.

in many respects related to this school. In some respects my views are close to theirs, in others, however, they differ considerably from those of the Clower–Barro–Grossman school. At the points of contact we shall repeatedly refer to the similarities and differences.

We have already reached the first point of contact: Clower emphatically drew attention to the *possible separation of intentions and realizations* when demand and supply were not in equilibrium. He asked what would happen in such cases. This was an important step forward in comparison with the traditional treatment of the problem, which always concentrated on the conditions for existence of equilibrium and the characterization of the equilibrium state.

Yet Clower himself only made *one* step forward. He suggested that if, for example, supply remains below demand, it is the "shorter side" that asserts itself: actual purchase is realized at the level of supply. That may be so, but it is only the beginning of the story, the continuation of which is no less important. What happens next? A whole series of events, involving forced adjustment: the buyer makes forced substitutions, he searches, revises his demand, waits, and so on. Clower grasps a *static* moment, namely the instant of half-success and half-failure when the buyer takes his purchases up to the supply constraint. In practice, however, the question is about the *dynamic processes* of instantaneous and lasting adjustments, which cannot be represented in a static model.

I would like to take this comparison further in a particular context: let us consider the role of forced substitution in the case of a conflict between intentions and their feasibility. For the purpose of comparison a strongly simplified example will do. We shall describe an individual buyer's behavior. He buys two substitutes, i and j.

Let us denote the buyer's initial demands by d_i and d_j, the supply by s_i and s_j, and actual purchases by y_i and y_j. According to Clower, to a first approximation actual purchases will be the following:[9]

$$y_i(t)=\begin{cases} d_i, & \text{if } d_i \overset{>}{=} s_i \quad \boxed{\text{no shortage}}, \\ s_i, & \text{if } d_i > s_i \quad \boxed{\text{shortage}} \end{cases} \tag{5.5}$$

[9]The remark "first approximation" has been added because both Clower and his followers discuss in detail secondary effects spilling over through spending or saving. (For example, the worker cannot sell his labor-power, so that because of his reduced income his demand for consumer goods is lower, etc.) The present book will also discuss indirect effects later, and comparisons will again be drawn with the views of the Clower–Barro–Grossman school. Yet now we shall disregard such spill-over. At this point we discuss the situation in which the *buyer's disposable income is given.*

i.e. $y_i(t) = \min(d_i, s_i)$. The determination of y_j is analogous to this. Formula (5.5) expresses the principle that it is always the "short side" that is asserted. In a later chapter I dispute this principle: in my opinion it does not sufficiently describe the real situation. However, for the sake of simplicity of the argument, I disregard this counterargument and temporarily adopt the assumption that "the short side is asserted".

For the moment I shall also disregard the fact that the process is dynamic: first the buyer perceives shortage, then he brings himself to make forced substitution, and so on. Let us therefore examine the final result – taken out of time – of the shopping process:

$$
y_i = \begin{cases}
d_i, \quad \text{if } d_i \leqq s_i \quad \text{and} \quad d_j \leqq s_j, \\
\boxed{\begin{array}{l}\text{no}\\ \text{shortage}\\ \text{of } i\end{array}} \qquad \boxed{\begin{array}{l}\text{no}\\ \text{shortage}\\ \text{of } j\end{array}} \\
d_i + \mu_{ji}(d_j - s_j), \quad \text{if} \quad d_j > s_j, \\
\boxed{\begin{array}{l}\text{initial}\\ \text{demand}\\ \text{for } i\end{array}} \ \boxed{\begin{array}{l}\text{forced}\\ \text{substitution}\\ \text{to make up}\\ \text{for shortage}\\ \text{of } j\end{array}} \qquad \boxed{\begin{array}{l}\text{shortage}\\ \text{of } j\end{array}} \\
\text{and} \quad d_i + \mu_{ji}(d_j - s_j) \leqq s_i, \\
\boxed{\begin{array}{l}\text{initial demand for } i\\ \text{and forced}\\ \text{substitution intention}\\ \text{can be satisfied}\end{array}} \\
s_i, \quad \text{if } d_i > s_i, \\
\boxed{\begin{array}{l}\text{shortage of}\\ i \text{ in relation}\\ \text{to initial}\\ \text{demand}\end{array}} \\
\text{or if } d_j > s_j \quad \text{and} \quad d_i + \mu_{ji}(d_j - s_j) > s_i,
\end{cases}
\qquad (5.6)
$$

$$
\boxed{\begin{array}{l}\text{shortage}\\ \text{of } j\end{array}} \qquad \boxed{\begin{array}{l}\text{initial demand for } i\\ \text{plus forced}\\ \text{substitution intention}\\ \text{cannot both be satisfied}\end{array}}
$$

or, in a more concise form,

$$y_i(t)=\min\left(s_i, d_i+\mu_{ji}(d_j-s_j)_+\right).$$

The determination of y_j is analogous with this.

The parameters μ_{ji} in (5.6) is the buyer's *forced substitution propensity* in the relation "j substitutes for i". Attention is called to the fact that this is *not* identical with the microeconomic concept "marginal rate of substitution". The latter expresses a voluntary substitution propensity: if both i and j are available on the supply side, how many units of j are considered the equivalent of a unit of i. In the situation discussed here the buyer decided this question earlier, and determined the vector $[d_i, d_j]$ of his initial demands accordingly. The parameter μ_{ji} indicates the following: *if* j is not available in the quantity d_j demanded initially, how much of i is the buyer willing to accept *in that case*? We do not ask whether, after buying this quantity, he considers himself compensated. Presumably he does not: he suffered a loss. We ask the following question strictly in the descriptive sense: if such a forced adjustment takes place, what will he do? This is, therefore, a *conditional* substitution rate under the conditions caused by shortage. It expresses how far the buyer is willing to go in *revising* his initial demand.

We shall come closer to an interpretation of μ_{ij} if we think of a sequence of recurrent purchases. Within this the buyer may hesitate with regard to forced substitution, bringing himself to undertake more on one occasion, and less on the next. The parameter μ_{ji} is the intertemporal average of this attitude variable.

Now let us consider each row of formula (5.6).

In the *upper row* the alternative in which there was no shortage of product i or j was described. Accordingly, actual purchases may take place in accordance with the initial demand.

In the *middle row* one of the alternative situations is shown, i.e. a shortage of j. To the initial demand, d_i, the buying intention entailed by forced substitution is added. Recognizing shortage, the buyer has revised his initial demand. In the case shown in the middle row this revised demand, including the addition due to forced substitution, can be satisfied. *Forced substitution absorbs not only excess demand but also excess supply or part of it.*

The *lower row* is divided into two subcases. One is "simple" shortage: initial demand for i is larger than supply. Therefore purchase takes place at the level of supply. The other is shortage brought about by forced substitution. Since there was a shortage of j the buyer tried to substitute i for it.

However, now the supply is not enough to meet the initial demand increased by the forced substitution demand; a shortage of *i* has also come about. And now the actual purchase of *i*, too, is constrained by the supply side, although supply may have been enough to cover the initial demand without forced substitution. (The sufficient condition in the last row does not exclude initial excess supply, $d_i < s_i$.) *With the appearance of intentions to undertake forced substitution shortage "overflows", that is spills over from the market of product j to that of product i.*[10] And if we turn from the simplified model of "two products, one aggregate buyer" to more complex reality, we shall understand this process even better. Buyers finding partial shortages on one or other partial markets move over to better supplied markets where they may cause further shortage; buyers finding shortage there again try to accomplish forced substitution and so on. And, while every buyer buys something (they make forced substitutions) nevertheless most of them are left with a "feeling of want".

The formula (5.6) is certainly more complicated than the basic formula (5.5) of the Clower–Barro–Grossman school. Yet, in my opinion, we must insist on (5.6). It helps to understand the seemingly paradoxical phenomenon of the shortage economy, that there is shortage of many products, and yet buyers spend the money intended for purchasing goods and services. Forced substitution (and other related forms of forced adjustment) is one of the keys to understanding the shortage economy.

We have to make one more terminological remark. Clower uses the expression "notional demand" to denote the purchasing intention prior to hitting a supply constraint. Since Clower describes a two-step shopping algorithm, the distinction between two categories is sufficient for him: those of notional and effective demand. Yet we consider shopping as a *process* in which the buyer can modify his demand more than once. This dynamic character of demand formation is indicated by the terms "initial demand", "demand revised for the first time", "demand revised for the second time", and so on.

Since Clower clearly defined the concepts relating to his two-step purchasing algorithm, I suspect that the best way to avoid terminological confusion is for us to introduce other terms for related – but not identical – concepts. We shall proceed similarly with a few other categories of the Clower–Barro–Grossman school.

[10] If the original buying or selling intentions hit a constraint on the other side, the effect may spill over from the market of the *i*th product to other markets. This spill-over effect is treated at length by the Clower–Barro–Grossman school, as mentioned already. (See, for example, Benassy, 1978, and Howitt, 1978.)

5.5. Observation and measurement of demand

The group of questions discussed in the previous section, concerned with the relation between intentions and realizations, is closely linked to our next subject: the observation and measurement of demand.

Let us consider the situation on a partial market on which only the ith product is traded. Let us assume that a total of m sellers sell the product, and they are confronted with n buyers.

We are at the beginning of day t. Let us assume, for the sake of simplicity, that the daily shipment is delivered to each selling point in the morning before the start of sales. The delivery of product i to selling point h is denoted $x_{ih}(t)$. The terminal stock $u_{ih}(t)$ is recorded in the evening when selling ends.

The kth buyer's initial demand for product i is $d_{ik}(t)$. The transaction of product i effected at selling point h with buyer k is denoted $y_{ihk}(t)$.

First, two extreme cases will be described.

In the first case the following condition will be satisfied:

$$\underset{\boxed{\text{terminal stock}}}{u_{ih}(t)} > 0, \quad \text{for every } h \text{ and } t. \tag{5.7}$$

If this condition is fulfilled, it clearly proves that the buyer was in a position to satisfy his initial demand every day at the selling point visited. If a positive stock remained on every occasion, and everywhere, it proves that more product could not be sold, there was no initial demand for any more.

Under such circumstances the following equality surely holds:

$$\underset{\boxed{\text{aggregate initial demand}}}{d_i(t)} = \underset{\boxed{\text{aggregate actual trade}}}{y(t)}, \quad \text{for every } t, \tag{5.8}$$

where

$$d_i(t) = \sum_{k=1}^{n} d_{ik}(t) \quad \text{and} \quad y_i(t) = \sum_{h=1}^{m} \sum_{k=1}^{n} y_{ihk}(t).$$

From the viewpoint of observation and measurement this means that it is sufficient to observe $y_i(t)$, that is the *actual* purchase, from which an inference can be drawn with regard to $d_i(t)$, the buyer's *intention*. The variable $y_i(t)$ is observable *objectively*, therefore it is not necessary to question the buyer about his intentions.

In what follows (5.8) will be called the condition for *objective observability* of demand.

In the second case the following condition is fulfilled:

$$u_{ih}(t)=0, \quad \text{for every every } h \text{ and } t. \tag{5.9}$$

Has initial demand been satisfied? *We do not know*. Fulfillment of condition (5.6) is logically compatible with each of the three subcases below.

(a) *There was no shortage* of product i and it was not used for forced substitution, either. Luckily, the morning delivery each day exactly covered the demand of that day, so that no stock remained by the end of the day:

$$\sum_{j=1}^{m} x_{ih}(t) = \sum_{h=1}^{m} \sum_{k=1}^{n} y_{ihk}(t) = \sum_{k=1}^{n} d_{ik}(t). \tag{5.10}$$

aggregate quantity of deliveries	aggregate actual trade	aggregate initial demand

(b) Although *initial* demand for product i could have been satisfied from the day's delivery, a further *revised* demand was put forward. Product i was purchased as a forced substitute for product j. This aggregate demand could then not be satisfied:

$$\sum_{k=1}^{n} \{d_{ik}(t) + \mu_{ji}[d_{jk}(t)-y_{jk}(t)]\} > \sum_{h=1}^{m} [u_{ih}(t-1)+x_{ih}(t)]. \tag{5.11}$$

aggregate initial demand for i	aggregate additional demand for i as forced substitute for g	aggregate supply

(c) The aggregate supply was too low even to satisfy the initial demand for i, independent of any intentions to buy it also as a forced substitute for another product in short supply:

$$\underbrace{\sum_{k=1}^{n} d_{ik}(t)}_{\text{aggregate initial demand for } i} > \underbrace{\sum_{j=1}^{m} \left[u_{ih}(t-1)+x_{ih}(t)\right]}_{\text{aggregate supply}}. \tag{5.12}$$

Now comparing the initial demand for i with the actual purchase of i we can see the following:

$$\begin{aligned} &\text{in subcase (a): } d_i(t)=y_i(t),\\ &\text{in subcase (b): } d_i(t)<y_i(t),\\ &\text{in subcase (c): } d_i(t)>y_i(t). \end{aligned} \tag{5.13}$$

From the above, the following *impossibility proposition* results.

When the terminal stock is always zero at every selling point, no inference can be drawn from the actual trade $y_i(t)$ (i.e. from the objective observation of realizations) about the initial demand $d_i(t)$, the original buying intention.

We have described two extreme "pure" cases. In reality in between cases are frequent: at one selling point there is a positive stock, at another there is not; at one time there is, at another time there is not. If a zero stock is a rare phenomenon, and at most places there is a positive terminal stock most of the time, the condition of objective observability of demand is fulfilled at least approximately. In this case it is permissible, as an approximation, to draw inferences from y to d. This is also permissible if the general depletion of stocks is only a temporary event. This is represented in fig. 5.1. Here the terminal stocks of the sellers are quite stable: this is represented by the dotted horizontal line. Although the stocks became depleted in period $[t_1, t_2]$, later they increased to their customary level. The time series of actual purchases also falls temporarily, but it, too, is soon restored to its customary level. In this case it is justified to assume that initial demand did not diminish, but remained unsatisfied through the fluctuation.

In the case of sporadic and temporary shortages it is a testable hypothesis whether the time series $y(t)$ reflect the demand or the supply constraint

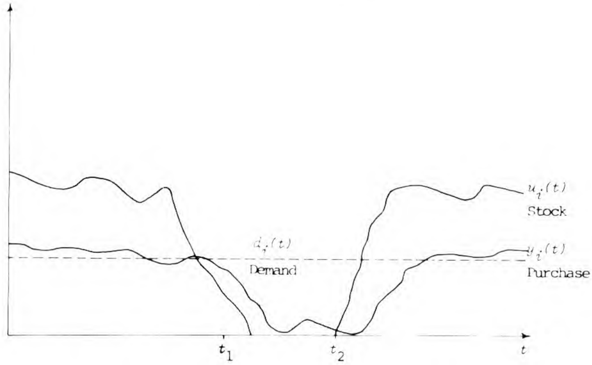


Figure 5.1. Demand at the time of temporary exhaustion of stocks.

in the given period. Inspired by the Clower–Barro–Grossman school, econometric methods have been developed to do this testing.[11]

The situation is radically different if shortages are not temporary but chronic, and not sporadic but widespread and frequent. Even if the preceding "second case" does not occur "purely", zero stock is quite a frequent event; hence the impossibility proposition becomes valid. *The inference from y to d becomes inadmissible.* And, this being a logical impossibility, no econometric technique can help it.[12] It is a vain effort to test whether the given purchase reflects initial demand satisfied, revised demand satisfied, or the supply constraint.

All this is not intended to convey that in a shortage economy nothing can be done to observe demand. A lot can be done – but it must be in the awareness that observation and measurement will not be objective, nor highly reliable. We may mention just a few of the possibilities for observation and measurement.

(1) The buyer may be *interviewed*, that is he can be questioned about his buying intentions and their fulfillment. This may be costly and tiresome, yet it is certainly very instructive. This is frequently done in market research. It is another question that the reliability of the buyers' answers show a rather large dispersion. In the final analysis no harm comes to the buyer from having given inaccurate answers to the market researcher, therefore many of them talk thoughtlessly.

[11]See, for example, Goldfeld–Quandt (1975).
[12]Such an attempt is made for example in Portes–Winter (1978).

Market research usually investigates demand for a single product or group of products. Answers refer to buying intentions directed at this one product, from the overall purchasing plan. Exactly on that account it may easily happen that this intention is inconsistent with the rest. It is questionable whether all the buyer's purchasing plans are in fact consistent with each other, and with his spending possibilities.

(2) There is a queue for certain products: in such cases the length of the queue is observable. Similarly there are products delivered exclusively to order: in such cases the backlog of unfilled orders, or the volume of rejected orders, can be observed.

Although these are important bits of information, it is doubtful that they can give a faithful picture of buying intentions. It may be that the buyer is maneuvering. For example, he orders the same material from three different places and if he receives it from one, he would cancel the order with the other two. He can always find an excuse for the cancellation. But even if he has to pay a penalty, in a shortage situation it is worthwhile augmenting chances by multiplying orders.

Because of all this, in the queues and in the backlogs of unfilled orders a *fictitious demand* exaggerated to facilitate maneuvering, is also present, in addition to the really "serious" demand. This is a phenomenon analogous to hoarding an input stock to give security, which will be discussed below. Orders are, in a certain sense, "negative stocks". The firm is interested not only in accumulating positive stocks of the inputs that are important for it, but also "negative stocks", that is a certain quantity of orders.

(3) In the transactions where no queue develops, the seller and buyer at least confer with each other. Let us recall the phenomenon of search. The buyer visits selling point 1 and does not find the required product. He tells the seller to report to those competent (producer, wholesale trade, superior directing authorities) that such and such a product is a "shortage good". The buyer's "voice", his complaint about shortage, is thus observable; for example, the seller may report on it. Inferences can be drawn from it about demand.

Yet distortions may also arise in this case. The buyer continues searching, visiting selling points 2 and 3, and so on. He complains to each one, and each seller reports on the shortage signals he perceived. And, the more patient the buyer is in search and the louder in complaining, the more this shortage signalling will be intensified. Ultimately, this may also lead to multiplication of perceived shortage, that is a phenomenon similar to "fictitious demand".

Here we can observe a vicious circle in the perception of shortage. The larger the real shortage, the more the buyer feels prompted to multiply the volume of his orders, to signal shortage at many places, augmenting fictitious demand. And the larger the fictitious demand, the more everybody "feels" shortage.

We must make one more remark of a theoretical character. Everything that has been said under items (1), (2), and (3) about the inconsistencies in buying intentions and about fictitious demand means at the same time that *buyers often violate* what is called the "Say principle" in modern mathematical economics.[13] This principle is understood to imply that the buyer's purchasing intentions are consistent with each other as well as with the buyer's financial abilities, that is they do not exceed the buyer's budget constraint. According to the above, the intentions *declared* by the buyer may be inconsistent with each other and already on that account realizations may differ from intentions.

5.6. Hoarding tendency

In sections 5.2–5.5 mainly questions of *methodology* and *history of economic thought* were discussed: how we should describe demand formation, how it can be observed and measured. It is true that these ideas were already suggested by the examination of the input demand of the socialist firm and in our above discussion a few remarks were made on the behavior of the firm. Now, in the present section, we shall leave methodological questions behind, and turn to *descriptive theory*. We wish to emphasize a few characteristic and empirically observable features of the formation of material demand of the socialist firm.

In the behavior of the socialist firm as buyer there is a *hoarding tendency*: the firm tries to accumulate as large an input stock as possible.[14] This tendency appears both in the traditional and in the postreform economic management systems. The firm is driven towards hoarding by a strong *quantity drive*. From the motives described in Chapter 3 the firm endeavors to increase production, and therefore needs more and more

[13]For interpretation of Say's principle I borrow the formulation of the paper by Clower (1965): it means the *ex ante* consistency of buying intentions and financial resources. Furthermore, see Leijonhufvud (1972). We return to the discussion of Say's principle in Chapter 13.

[14]We shall see later that a similar tendency is also present with other inputs. "Labor hoarding" within the firm resembles material hoarding.

input. This tendency is further strengthened by uncertainty in the sphere of production and trade which has just been discussed.

There are, however, certain countertendencies opposed to the hoarding tendency; its assertion is impeded by various constraints.

(1) The material purchased must be stored and *storage capacity is limited*. It is true that it could be increased through investments, but investment resources are needed for a great many other purposes as well. However trivial this constraint may seem, it is perhaps the most unambiguous one. A "clever" firm will pack its store-rooms if it can, but it cannot hoard more than that.[15]

(2) Superior authorities use various methods to put pressure on the firm so that it should not hoard too much material. A role may be played in this by the system of material rationing, which will be discussed at the end of the chapter. Also, the Central Bank can restrain credit for financing material stocks. This factor has strengthened in Hungary since the reform.

Beyond administrative and financial measures, "moral pressure" is also exercised. The firm engaging in excessive material hoarding may be criticized at meetings or in the Press.

(3) The firm may voluntarily, using its own judgement, restrain its material demand.

The three factors enumerated above put tolerance limits or acceptance constraints on the hoarding tendency. Primarily, they exert their effect in the direction of restraining the firm's *total* input stock (or perhaps the total of input *and* output stock). The storage capacity restricts the total stock. The practice of granting credit is also concerned with aggregate stock data. Therefore, the material hoarding tendency of the firm appears in a *selective form*. It tries to accumulate as much as possible of the essential or hard to substitute materials or those where the supply position is uncertain, and it is content with smaller stocks of easily substitutable or easily available materials.

In the firm's behavior two seemingly contradictory phenomena are present. On the one hand, instantaneous initial demand at the submicro-level is always determined and finite. On the other hand, demand is – aggregating over inputs and over firms and for a long period – insatiable and thus tends towards infinity.

Instantaneous initial demand is *determined*, because it is developed by the rules described in section 5.2. And yet, in a certain sense, demand is *for*

[15]Limited storage capacity does not stop the buyer, of course, from placing as large and as many orders as possible.

ever unsatisfied. Given the uncertainties of production and trade the firm not bound by an effective budget constraint, living in an atmosphere of quantity drive, will never say: "I have enough material. No matter what material I could get more of, I shall voluntarily refuse purchase." And, as long as it does not make such a statement, it can be said that its material demand is insatiable.

This peculiar double state will be referred to in what follows by saying that the firm's demand for material inputs for current production is *almost insatiable*. The word "almost" indicates that demand is not "really" infinite, but only that a tendency towards infinity dominates it, impeded somewhat by countertendencies and constraints.

The hoarding tendency and, together with this, the almost-insatiable demand for inputs for current production represent one of the fundamental processes that "siphon off" external slack (in this case the seller's output stock) from the economy. In this way they intensively promote permanent reproduction of the shortage state of the economy.

We have again arrived at one of the vicious circles of shortage phenomena. The firm and its superior authorities perceive the almost-insatiable demand for their products. This intensifies the quantity drive. Given the uncertainties of production and trade a hoarding tendency appears. Such a tendency in turn amplifies material shortage. And, the more intensive the shortage of some material becomes, the more that material will be hoarded, and the more uncertain its supply becomes – then the quantity drive will be pursued most intensively in the production of exactly these goods. *Thus quantity drive, hoarding tendency, and material shortage are linked together into a mutually reinforcing, self-sustaining process.*

This train of thought can also be reversed. If there were no material shortages repeatedly threatening, the firm would not worry about its future material supply. It would be confident that the seller would always have the required output stock, or he would be prepared to adjust his production quickly to the needs of the buyer and so the firm would not strive to accumulate input stocks in its own warehouses. And that is exactly what would discourage the appearance of any shortage.

Discussing the accumulation of input stocks, one further important distinction must be introduced. In practical economic life a distinction is made between "frozen" and "current" stocks. The former category includes stocks that have been in the possession of the firm for a long time and there is hardly any chance of their utilization within a reasonable time. As a matter of fact, it would be more exact to introduce a continuous scale to describe the "frozen-current" character of the stock. For our purpose, however, a rougher distinction seems sufficient. Since, however, we wish to apply this classification to both the input and the output sides, and for products as well as for services, we introduce the pair of concepts, *productive and unproductive slack*. A special case, for storable goods is the "frozen" and "current" stock distinction mentioned above.*

We wish to emphasize that these are *ex ante* categories. Slack is unproductive if there is no (or negligibly small) chance of its utilization within a reasonable time horizon, either for its original purpose or in a closely related forced substitution.

Recall that in section 2.6 we distinguished between the two different states of the internal slack of the producer firm: the mobilizable and non-mobilizable slack. There were categories of instantaneous adjustment. Slack is not mobilizable if at the moment of utilization one of the inputs needed for its utilization is missing. As opposed to this, the concepts just introduced, productive and unproductive slack, are categories of *short-term* adjustment. Some input stock becomes unproductive if it is not mobilizable in the foreseeable future (because of the shortage of complementary inputs), or perhaps it is not even needed (e.g. because the output composition has changed and future output will no longer use that input).

If one element of slack is described as productive, this does not mean we always "approve" of its acquisition. We do not say that its purchase is expedient and advisable. Raising this question actually goes beyond the subject of the book and leads to operations research, dealing with the problem of "optimum inventory" and "optimum reserve". The attribute "productive" only asserts that there is reasonable chance of a proper utilization of the slack in question.

Since the qualification "productive–unproductive" is based on *expectations*, it may change if the decision-makers' expectations change. The purchasing agent of the screw factory bought 100 tons of steel of quality 7 at the end of 1976, because he thought it would be needed soon. Two years have passed since. It has never been needed, and now it is clear that it will not be, either. These 100 tons have become part of the unproductive slack.

And here we come back to the problem of hoarding. Shortage sometimes leads to *buying hysteria*. "I shall buy it, it may be good for something." It is all the easier to think in this way when money does not much constrain the buyer. Then part of the recklessly bought materials, semifinished goods and parts become frozen stock, that is unproductive slack. This may even

*The terminology *active* and *inactive* stock may also be found in the Western literature, but our use of the terms productive and unproductive slack is perhaps more accurate. (*Editor's note.*)

add to the shortage of inputs elsewhere, for example exactly where that material might be needed.

The existence of this hoarding tendency is a hypothesis which is *empirically testable*. (We shall revert to this in the next chapter.)

The hoarding tendency accounts for the fact that the level of input stocks considered *desirable* by the firm is often higher than its *normal* level developing as an intertemporal average. The *desired* stock which appeared in formulae (5.1)–(5.4) is the buyer's *aspiration level*. He would like to achieve it, *ex ante*, when drawing up his shopping list. As opposed to this, the *normal* stock is what develops in conditions of chronic shortage, given supply constraints, *ex post*, on average over a long period. The actual stock fluctuates around the normal stock. The hoarding tendency and, as its expression, the high *ex ante* desired level always drives upwards the *ex post* normal level of the input stock.

5.7. Central material rationing

So far in this chapter, and in Chapter 4, the subject under discussion has been the nature of the linkages between buyer and seller. In fact, these linkages might be considerably influenced in a socialist economy by *central material rationing*. This has already been mentioned in section 4.1; what was said there must be repeated here for the sake of completeness.

Various administrative authorities, such as the Central Material Rationing Office or ministries directly controlling production draw up material balances and indicate the limits of transactions: from where and to where the main flows of material inputs should go. Central material rationing assumes an important role in the information gathering activity of buyer and seller and at the same time restricts their freedom of decision.

It restricts it – but in no way eliminates it. Let us examine in somewhat more detail the situation of firms functioning in the *traditional* economic management system. (Later we shall briefly discuss the postreform situation.) On an abstract level the absolute centralization of material rationing is conceivable: it would be centrally decided down to the smallest detail which output should go where and when. Yet reality is always different: even in historical periods when centralization went farthest, multilevel control was going on, and within this framework the firm still had a sphere of movement. This was asserted in two principal forms.

One form is where *the firm exerts an influence on the central decision* prior to decision-making. In most cases the firm has to submit its *claim* for the coming short-term plan period before the Material Rationing Office makes its decision. And, even if the material rationing system functioning in the country in question does not prescribe the submission of formal claims, this is done informally in any case. Through personal discussions, letters, or by using the mediation of other superior authorities or political organs, the firm will let the Material Rationing Office know about the kind and volume of material it wishes to receive.

Claims (formal and/or informal) and *demands* are related phenomena. Their relation is demonstrated in fig. 5.2. The direction of a claim is *vertical*: it is directed to the superior authority; the direction of a demand is *horizontal*: it is directed to a selling firm legally on the same level. Yet in practice both appear in a relation in which *the buyer firm is in a subordinate position*. It is dependent upon the central authority in the administrative sense and also on the selling firm in the economic sense, since there is a "seller's market". This similarity in the situation, in the buyer's position, accounts for the numerous points of similarity in the formation and assertion of claims and demands.

There is usually a demand, that is a buying intention behind a claim. And it is a demand of which it was said earlier that it is almost insatiable. In Chapter 3 we have already discussed "plan bargaining" in connection with "tautness" of the output plan. The other side of this same plan bargaining process appears in connection with material rationing. The firm can get along smoothly and fulfill its plan more securely if with given input quotas smaller output targets are prescribed and, conversely, if with a given output target larger input quotas are allotted. Therefore the firm tries to "jack up" its material claim. It knows that only 200 tons are necessary,

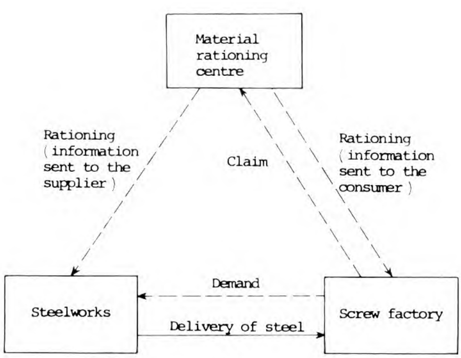


Figure 5.2. Claim and demand.

but claims 220 instead. But the Central Material Rationing Authority also knows the rules of the game and "cuts down" the claim to 200 and perhaps even to 195.

When discussing the "insatiable nature" of material demand we pointed out that various factors do set upper tolerance limits. These also affect claims; and particularly strong are "tactical" considerations. The experience and detailed information held by the Material Rationing Authority must be taken into account. If the firm asks for 220 tons, although in fact it only expects 200, this is presumably part of the "game". The 20 tons are really subject to discussion; serious arguments may be brought up. If, however, the firm asked for 300 tons, that would harm its credibility. A special kind of "self-censorship" is functioning: it is not worth making any claim that would be flatly refused in any case. The "plan bargaining" process will repeat itself again and again and it is in the firm's interest not to worsen its position in these bargains in the long run. This may be the most important factor setting an upper limit to the tendency to exaggerate needs in making claims.

The firm, with its claim, is "competing" with other claimants for the scarce material allocations. Therefore it strives to gain the favor of the rationing authority. In this connection many of the same phenomena arise here as were treated in section 4.6, where efforts to win over the seller were discussed.

The other form in which the sphere of movement of the buyer and seller firms can be asserted is *a certain amount of freedom of action in executing central decisions and in transactions not controlled centrally.*

Central material rationing is usually specified in a relatively aggregated form. "The screw factory may buy 300 tons of steel of quality 8–12 in the second quarter." There is no closer indication of quality, and no time schedule. For certain periods and certain materials it was also not indicated from which supplier the material had to be purchased. The Central Rationing Office often just sanctions the agreement signed between seller and buyer. *Legally* every allotment must be completed by means of a *"business" contract* between the buyer and seller firms.

What is more, not even in the most strongly centralized periods was central material rationing extended without exception to every material and to every particular transaction. A fairly wide range of materials remained outside the sphere of central material rationing.

Therefore, *central material rationing is a mechanism which considerably influences the input–output flows between firms but does not eliminate the main features of the relationship between seller and buyer, its characteristic* *intrinsic regularities*. All that was said about these features and regularities earlier, and to be explained in the following chapters, is also valid and effective in the circumstances of central material rationing.

Central material rationing uses various means to prevent the hoarding tendency. The tendency asserts itself in spite of that, and partly even because of it. The firm is uncertain not only about whether – in the real sphere – it can "physically" obtain the required input. It is no less uncertain whether – in the control sphere – it can obtain sufficient material rations from the authorities.

Horizontal shortage perceived by the firm in its role as buyer is complemented (or even preceded) by *vertical shortage* which the firm experiences as an applicant for rationing. Horizontal shortage is the shortfall of the actual purchases from the instantaneous buying intention. Vertical shortage is the shortfall of the actual ration from that applied for.

And now let us comment on the postreform situation. Since the reform, central material rationing has been almost completely abolished: it now covers only a few materials. This removed many of the difficulties caused by the intricacy and rigidity of material rationing. Yet it did not eliminate the problems caused by the resource-constrained character of the economy and by the chronic reproduction of shortage.[16]

The question of how far the formation of the firm's demand, its hoarding tendency, and other phenomena mentioned above are affected by profitability, credit, and interest in the postreform economic management system, will be discussed in Part II of the book. At this point we can only suggest that since central material rationing now has a very narrow scope, the phenomena described in this chapter and the last exert all the greater influence. The intrinsic regularities of horizontal relationships between buyer and seller assert themselves.

[16]Part of the inputs intended originally to be bought from domestic producers is quickly substitutable by imports in the case of shortage. This relieves shortage instantaneously. At the same time, in such cases a special "reallocation" of shortage takes place: it is shifted from the sphere of domestic trade to a burden on the international balance of payments. This "reallocation" of shortage will be treated further in Chapter 21.

CHAPTER 6

The seller

6.1. Introduction

Neoclassical microeconomics accustoms us to regarding the roles played by buyer and seller as perfectly symmetrical. The buyer has a demand function, the seller a supply function, and both depend on prices. With an increase in price, the buyer responds by reducing demand and the seller by increasing supply.

This symmetry is, however, only complete if both buyer and seller have a hard budget constraint. It is then very important for the buyer to spend disposable money in a reasonable way, and he therefore reacts sensitively to the purchase price. And it is very important for the seller to receive financial income, and he therefore also reacts sensitively to the selling price. Of course, there is also a basic asymmetry[1] under such circumstances: the seller supplies a physical product, that is *real* goods, and the buyer gives *money* in exchange. Yet money is a "serious matter": if he does not buy the goods he is negotiating about with the seller, he can buy something else for it. To give money for a physical product requires as much consideration and responsible decision as to give away a physical product for money.

It is precisely money that allows any concession or extra on the one hand to be counterbalanced by a concession or extra on the other. If the buyer makes some allowance in regard to the physical attributes of the product (e.g. he accepts a poorer quality), the seller may grant a discount on the price in exchange. And, conversely, if the seller gives an extra service (e.g. he delivers goods to this final destination), the buyer can also give something extra in exchange, by paying a surcharge. I do not claim that this "equivalence" is always asserted, only that in the case of a

[1]In this connection Marx mentions the "salto mortale" of the commodities. (See Marx, 1867–1894b, *The Capital*, vol. I, p. 106.)

"serious" functioning of money and prices any *physical* plus or minus can be counterbalanced by some *monetary* plus or minus.

The situation is essentially different if money does not play a really "serious" role; if it does not actively influence decisions but just takes part in the passive recording of the process. The economic unit whose budget constraint is soft, that is its permanent survival is guaranteed even in the case of a lasting financial deficit, will respond less to prices. And in this case the "counterbalancing" role of money is weakened. For example, the seller lowers the quality of the product, but this may not even be reflected in the selling price. And if it were; and if legal rules required a discount? The seller would not mind much; ultimately, it is not a matter of life and death how much the financial income amounts to. And the reverse is also true. The buyer would not feel compensated, either, if he had to pay less for a product of poorer quality. It is true that his costs would be reduced in a certain sense, but, in the final account, this does not really matter at all.

The weak influence of relative prices already implies the *possibility* of asymmetry, because the price cannot play its "counterbalancing" role. The asymmetry is then *brought about* by shortage. (Which again is closely connected, as suggested by the whole thought process of this book, with the weak role of prices and money.)

On one hand there is the buyer with his almost-insatiable purchasing *intention* – and on the other there is the seller, holding the products which provide the *possibility of realizing* this intention. On the buyer's side a "mental" phenomenon is present (buying intention); on the seller's side a *physical* phenomenon (the product actually available). The buyer sets out, with his initial demand *in mind*, on his shopping route. Meanwhile he is not interested in what is in the seller's mind, but what he has in *stock*.

The seller feels that he supplies "something serious" to the buyer, that is a tangible and usable product. What he receives for it (the selling price, i.e. money) he does not consider really "serious". Therefore, in practice he does the buyer *a favor* in complying with his requirements. And it is not only the seller who feels in this way, but the buyer, too. Exactly because his budget constraint is not very hard, he does not think it very important how much money he spends on inputs. As a matter of fact, he is grateful for every favorable gesture on the part of the seller. *They are not on equal terms*, which they both recognize, and behave accordingly.

In my book, an attempt is made to understand both symmetry and asymmetry in the roles of seller and buyer. Among other things I also dispute in this regard those theories claiming that there is perfect symmetry between "excess demand" and "excess supply", and between "suppressed inflation" and "suppressed deflation", and so on.[2] While there are, doubtlessly, symmetrical features in these contrasting pairs, they also have numerous asymmetrical features. If shortage is general and chronic, if the behavior of micro-organizations has been adjusted to shortage, and if the role of money and price has weakened, the system shows a number of specific regularities which are *not* symmetrical contrasts to the regularities of an economy with low resource utilization and "Keynesian unemployment".

In general, theorists are fond of symmetry. We suffer from the sensation of an "aesthetic loss" if two phenomena thought of as strictly symmetrical turn out not to be so. I think, however, that the desire to provide a more faithful description of reality may compel us to abandon some of the more contrived symmetries.

6.2. Long-term, short-term, and instantaneous supply

The ideas about symmetry and asymmetry between the roles of buyer and seller were advanced in order to elucidate the discussion that follows on the seller's behavior, and supply. And now let us begin our examination of the subject of this chapter: selling. In a few respects, discussion must be narrowed down.

(1) In the two preceding chapters it was left open whether the seller – facing the buyer firm – was a producer firm or a trading organization. In the present chapter, however, we deal exclusively with the selling activity of the *producer* firm. Because of limited space, we will not treat the selling functions of the commercial firm, or trading organization.

(2) As in the two preceding chapters, we again discuss exclusively the trade in storable *products*, and not that concerned with services.

(3) The firm may, as we mentioned earlier, produce to stock or to order. For the moment, let us restrict ourselves to an examination of production *to stock*. The alternatives – production for stock or to order – will be discussed in section 6.3.

First of all let us consider the three degrees of adjustment and control. We start at the *general* level. For the moment it will be left open whether the firm in question is a capitalist or a socialist one.

[2]In the formulation of Barro and Grossman: "The main point of the present paper is that the existence of suppressed inflation has consequences which are completely analogous to the recognized consequences of suppressed deflation..." (see Barro-Grossman, 1974).

In the formation of *long-term* plans and investment decisions the firm's roles as producer and seller are not separated. The firm (and, in the case of multilevel control, its superior authority) expects that it will be able to sell what it manufactures. "Long-term production plan" and "long-term supply" are synonymous concepts.

In the elaboration of *short-term* (of yearly or even more of quarterly) plans the two roles may already be separated. (They *may* but, as we shall soon see, they need not.) If the firm regularly holds a substantial output stock, formation of the production intention and the selling intention require separate decisions. Let us reserve the concept of "short-term supply" to denote the short-term *selling* intention which may more or less coincide with but may also differ from the short-term production plan.

Finally, in the sphere of *instantaneous* adjustment (remaining with the case of a firm producing to stock) the *selling process* of the seller may appear. (This is analogous to the purchasing process described in detail in Chapter 4.) The seller has an *initial supply* which he can modify according to need on one or several occasions. Of course, behind the instantaneous selling intention there is physical supply, that is the output stock accumulated in the warehouse as "cover".

Thus "supply" is a collective term indicating a wide range of phenomena of which we have distinguished three specific kinds. Also, three kinds of "supply functions" can be mentioned according to whether the dependent variable of the function is long-term, short-term, or instantaneous supply.

After this conceptual clarification at a general level, we shall now proceed to discuss the firm functioning within the *resource-constrained socialist economy*. (At this point we need not specify whether the firm operates in the traditional or in the postreform economic management system.)

Long-term control is treated later (mainly in Chapters 9, 10, and 14); therefore, let us leave it aside for the moment.

Short-term control of production and selling can hardly be separated. The more intensive the shortage of the firm's products, the more certain it can be that what it has once manufactured it can also sell. A considerable part of the output – almost "hot" from the last finishing operations – is accepted by the impatient buyers.

In neoclassical microeconomics it has become a widely held view that production and selling intentions are self-evidently identical. Such a view is held without reservation, and applied to every system. And yet this is valid in fact only for a resource-constrained economy and not for a demand-constrained economy; though it is actually rather the latter that is kept in mind in neoclassical microeconomics in constructing its models.

Short-term planning of *production* was analyzed in Chapter 2 and what was said there can be extended in the sense of the preceding – to cover the short-term formation of *selling* intentions as well. We have shown how the volume and composition of production are affected by shortage signals. This feedback control mechanism may be called the *short-term supply function* of a resource-constrained economy. In section 6.5 we shall deal with it in more detail.

Let us consider the last degree, that is *instantaneous* adjustment. There the asymmetry between seller and buyer, of which we talked in the previous section, is clearly revealed. Table 6.1 draws a comparison between the pure resource-constrained and the pure demand-constrained system. (In reality, of course, systems are different from the pure cases of theory; therefore, the states represented in the table do not assert themselves always and everywhere in such extreme forms.) Neoclassical Walrasian microeconomics connects the upper right and lower left-hand corners of the table: buyer and seller are both able to accomplish the transaction by a single-stage action of purchase and selling. In reality, for at least one of the two parties, purchase or selling requires a longer process. In the resource-constrained economy under close examination here, it is the buyer who

Table 6.1
Process of buying and selling in the resource-constrained and demand-constrained system.

	Resource-constrained system (suction)	Demand-constrained system (pressure)
Buying	Initial buying intention may have to be revised several times	Buying intention may be realized at once
	Initial or revised buying intention may hit supply constraint	Buying intention does not hit supply constraint
	Buying process requiring a longer period	"Buying process" is reduced to a single buying action
Selling	Selling intention may be realized at once	Initial selling intention may have to be revised several times
	Selling intention does not hit demand constraint	Initial or revised selling intention may hit demand constraint
	"Selling process" may be reduced to a single selling action	Selling process requiring a longer period

Table 6.2
Temporal aspects of demand and supply.

	Instantaneous demand			Instantaneous supply
	Buyer A	Buyer B	Buyer C	
Monday	Initial demand	Demand revised once	Demand revised twice	Initial stock of Monday morning + Monday's production − Monday's sale
Tuesday	Demand revised once	Demand revised twice	Purchase	Initial stock of Tuesday morning + Tuesday's production − Tuesday's sale
Wednesday	Demand revised twice	Demand revised for the third time	Demand not yet made	Initial stock of Wednesday morning + Wednesday's production − Wednesday's sale

requires a longer shopping process, as we could see in Chapter 4. On the other hand, the seller may expect that his instantaneous supply (or at least most of it) will be sold without any particular difficulty. Therefore, we shall not be oversimplifying the matter if in the following we simply take as identical the instantaneous selling intention and instantaneous output stock, that is *physical* supply.[3]

In a suction economy, instantaneous supply (understood as the instantaneous output stock) encounters instantaneous demand that has reached various degrees of maturity. Some of the buyers come forward with their still fresh initial demand, others already have their first disappointments behind them and have modified their demand, perhaps already several times. This is shown in table 6.2 as a continuation of the left-hand column of table 6.1.

The output stock lying in the finished goods store of the producer-seller firm is one element of the slack of the national economy. From the point of view of the consumer-buyer firm this is an *external* slack, as opposed to the input stock accumulated in its own store, which is an *internal* slack.[4]

[3]Instantaneous selling intention is a variable of the control sphere, and output stock is a variable of the real sphere. While a close relationship exists between them, they are still *two kinds* of category.

[4]The distinction "external versus internal" is defined here, as well as at other points of the book, from the *user's* point of view. Use of internal slack is reserved for the owner, while external slack is available in principle to anyone discovering it who is willing to pay for it.

The expression "slack" is used here in the same spirit as earlier, for example in Chapter 2. It does not imply any value judgement. Slack appearing in the form of an output stock is in itself neither "good" nor "bad".

On the one hand slack helps to accommodate the differences in the timing of production and consumption. It serves as a "buffer" in the case of an unexpected increase in the buyer's demand. In this sense there are hardly any "unwanted" output stocks. Even a stock which appears almost superfluous may sometimes serve as a reserve, at least for forced substitution.

On the other hand, the same output stock is undoubtedly a "loss", an idle resource. With perfect coordination of production and consumption, unerring foresight and the elimination of unexpected problems an output stock would hardly be necessary.

6.3. Production for stock and to order

In the previous section only the case when the firm produces for stock was discussed. Here the question arises, on what does the producer-seller firm's decision to hold output stocks depend? And, if it does hold stocks what size will they be? Together with this, a further question arises, namely on what does the producer-seller firm's decision to accumulate a backlog of unfilled orders depend? And how large would it be? The following questions are more or less equivalent to the ones above: is there any queuing for the product, and if so, how long is the queue? The backlog of unfilled orders is, in an abstract sense, a "negative output stock". In the case of a positive output stock the seller waits for the buyer, in the case of a "negative" one the buyer waits for the seller.

Bearing in mind the primary subject of the book the answer would seem to be obvious, that it all depends on how intensive the shortage is, or, in the reverse case, how serious the sales difficulties are. It is true that all this *also* depends on the fact of shortage or sales difficulty. Let us, however, first consider the factors that are *not* closely connected to shortage.

The more differentiated the need fulfilled by the producer, and the larger the share of more expensive and indivisible units in the firm's output, the more useful it is to wait until the buyer announces his exact requirements. It is not only in shortage economies but also in "low employment" systems struggling with sales difficulties that ships or hydroelectric generators are not manufactured for stock; it is only after the detailed specification of the buyer's requirements that manufacturing begins. On the other hand, rails and reinforcing steel are produced for stock.

The borderline is not permanent. When cars were first manufactured every item was made individually. Later it was precisely in the car industry that the assembly line first appeared, and cars were produced in increasingly long series for stock. Yet more recently, customers' requirements have become increasingly differentiated in the industrially advanced countries. The car has a number of fittings, from radio to air-conditioning, that can be built into the car. The color of the car-body, the seats and even of the window-panes may be varied. The total possible combinations of all the optional features of the car is over a million. It is impossible to keep every variation in stock. Therefore, the larger car factories increasingly adopt the method whereby the buyer no longer buys the car from stock but orders it in advance. Then every detail of the manufacturing goes into the computer's program, according to the buyer's wishes. In this case production for stock has been replaced, at least partly, by production to order.

As is indicated by the example of the car, *technical progress* influences the proportions between manufacturing for stock and to order. Tendencies are not uniform; they develop in different directions for different products. From the phenomena connected with technical progress we only mention the following.

(1) As in the instance of cars, *needs differentiate*, which orientates producers towards the form of production to order. Yet there is also a countertendency, namely *standardization*. Reverting to an earlier example, screws are not manufactured to individual order; their sizes are standardized. And this indicates that production is for stock.

(2) *Production* techniques are changing. For example, building has traditionally been done to individual order. Since, however, building from prefabricated elements was invented, it has become possible to keep these in stock.

(3) *Storage* techniques themselves are changing. For example, the computerization of the storage of machine parts allows sales from stock on a much larger scale than at the time of less well organized warehouses when it was more convenient to order parts individually.

If we are now in a sector in which the circumstances mentioned above justify sales from stock, the question of how large the stock ought to be is still open. To a certain extent this is also independent of the shortage situation. Most products do not come into stock continuously, but "in batches". And nor do buyers wait for products to be delivered, but submit their claims periodically. Therefore, it is inevitable that there should be frequent *lags* between the arrival of goods and their forwarding to the buyer. This necessarily creates a transitory output stock. The less the arrivals and forwarding are coordinated, the larger this transitory output stock will be.

Let us summarily refer to all the factors listed so far as *organizational factors* influencing the proportions and size of the output stock and of the backlog of orders. A sharp contrast must be made between the influence of these factors and that of *shortage* or *sales difficulty*. If the output stocks of the screw factory are sold out and unfilled orders have accumulated in the office, this is not caused by organizational factors but by shortage. The same is true for the reverse case: if the factory has an output stock of six months' or a year's production lying in its warehouse, this is again not caused by organizational factors, but it is sales that have met with difficulties.

6.4. Normal input and output stock, normal backlog of unfilled orders

In this and the previous chapter, we have discussed in detail the factors which influence the size of input and output stocks as well as the backlog of unfilled orders. Some of them are *organizational* factors, others are related to the *resource-constrained or demand-constrained nature, that is the state of suction or pressure in the economy*. As a result of the persistent operation of all these factors *norms* develop. An economy adjusts itself both to the usual influence of organizational factors, and to the accustomed degree of shortage (or sales difficulty). In a particular country, and for any given range of products what is considered to be the *normal input and output stock*, or the *normal backlog of unfilled orders* is gradually established. Related to the latter is the *normal length of the queue* or *the normal waiting time*.

These norms as well as others involved in the process of controlling the economy are historical formations. They may increase or diminish if a permanent change takes place in the organizational conditions or in the distribution of market power. Yet at any instant these norms are given for the participants in economic processes. Summarizing what has been said so far we shall set up four empirically testable hypotheses.

(1) *In a resource-constrained economy normal input stocks form a relatively large share of total normal stocks, while normal output stocks form a*

relatively small share;[5] *in a demand-constrained economy the situation is reversed.*

This is a suggestive new example of the asymmetry – already mentioned several times – between the situations of buyer and seller.

Because of chronic shortage the firm as buyer strives to accumulate input stocks. (This was discussed in detail in the preceding chapter.) At the same time, it experiences no serious difficulties as seller; buyers suffering from chronic shortage absorb the greater part of the output stocks.

On the other hand, in the demand-constrained system sales difficulties lead to rising output stocks. It is worthwhile for the seller to protect himself against the uncertainties of sales by holding large output stocks: then if the buyer calls on him, he would not leave empty handed. In the buyer's role, however, it is less worthwhile to accumulate an input stock beyond the level necessary for organizational reasons. It is cheaper and more convenient for him to rely on the output stocks of the seller. Purchases cause him no particular difficulty.

In a suction economy every economic unit tries first of all to develop its own *internal* slack. "I can only be sure of what is in my store." In a pressure economy they are more prepared to rely on *external* slack. They think the owner of the external slack will exert himself to deliver inputs to them, the users.

To formulate the second hypothesis, output stocks must be examined in a disaggregated form. Even at a high intensity of shortage there are some products which the buyer does not want to accept even for a forced substitution, either because complementary inputs necessary for their utilization are not available, or because the quality of the products is so poor that they are unacceptable even to meet lower requirements. A frozen output stock consisting of such products is described – according to the classification introduced in the preceding chapter – as unproductive slack.

Of course, as was also indicated in the previous chapter, the delimitation is not sharp, since the more intensive the shortage the more the buyer is

[5]This is proved convincingly by the works by Farkas (1976) and Chikán–Fábri–Nagy (1978).

"In Hungary about two-thirds of stocks are held by users (buyers, further processors), and less than one-third by sellers (producers and wholesale dealers). International practice is just the reverse. Most stocks are held by sellers and only a small part by buyers. If, in addition, we take into account that a considerable part of both producers' and users' stock is made up of frozen stocks, valueless from the point of view of material supply, that provides an explanation of why service from store is irregular in Hungary, and why the transit time of orders and deliveries is long" – State Secretary of Heavy Industry, Adám Juhász, writes in his newspaper article (Juhász, 1978).

forced to accept inputs that are not to his liking. He may even be expressly instructed by the Central Material Rationing Authorities to do so. Notwithstanding this, for the sake of simplification we stick to our classification into productive and unproductive (current and frozen) output stocks.

(2) *In a resource-constrained economy the normal level of productive output stock develops in the neighborhood of the minimum level justified by organizational factors. The normal level of unproductive (frozen) stock considerably exceeds the minimum level justified by organizational factors.*

The first part of the hypothesis follows logically from the same arguments which were used to justify the first hypothesis: the buyer suffering from chronic shortage, with his almost-insatiable demand, and prone to hoarding anyway, will siphon off productive output stocks from the seller's store as quickly as he can.

The second part of the hypothesis throws light on the fact that in reality even the system that is basically resource-constrained is not "purely" so. This was also pointed out in Chapter 2: the attribute "resource-constrained" refers to the stochastically revealed *main* property of the system. Although rarely, it does happen that in one or other firm, with respect to one or other product, the demand constraint becomes effective even in the process of instantaneous adjustment. These exceptional cases, however, prove the rule, that in most cases the instantaneous effective constraint on an increase in output is the availability of resources.

(3) *In a resource-constrained economy the normal backlog of unfilled orders considerably surpasses the minimum level*[6] *justified by organizational factors.*

The three hypotheses mentioned above throw light on a new aspect of the phenomenon that has been discussed in the book several times: that shortage and slack are simultaneously present in a resource-constrained economy. This is not only because, as a result of organizational factors, stocks *must* accumulate even amid the most intensive shortage (there is inevitably a delay between production and use, etc.). The truth is that shortage and slack do not simply *co-exist*, but there are numerous *causal*

[6]The situation is well characterized by an interview given by the General Manager of "Merkur", the Hungarian firm entrusted with the exclusive right of importing cars:

Question: "In your opinion, what would be the ideal waiting time?"

Answer: "The car represents a high value. In order that the company can draw up plans and make decisions an adequate backlog of unfilled orders is needed. I would consider acceptable a period of one to one and a half years and I hope to achieve this at some later date, though, unfortunately, not in the near future."

See Moldován (1977).

relationships between them. Our hypotheses add some new points to the explanation of these relationships.

Shortage stimulates the firm as buyer to accumulate input stocks (internal slack). Development of frozen output slack (unproductive external slack) draws away resources and thus, indirectly, it may amplify shortage. The firm in the roles of buyer and producer does not take care not to accumulate an unproductive output stock, because it expects that it will be absorbed sooner or later, as a consequence of shortage.

Finally, our last hypothesis draws a comparison between the consequences of the *traditional* economic management system of socialist economy and those of the *postreform* system.

(4) *In the postreform economic management system the share of normal input stocks within the total stock is reduced as compared with the situation in the traditional system; furthermore, the normal share of unproductive output stocks within the total output stock is also reduced. This indicates that the intensity of shortage lessened, though it still remained quite considerable after the reform.*

Our hypotheses – if they are duly verified empirically – are statements of a *descriptive* character. In addition, however, a few *methodological* lessons can also be drawn from their formulation.

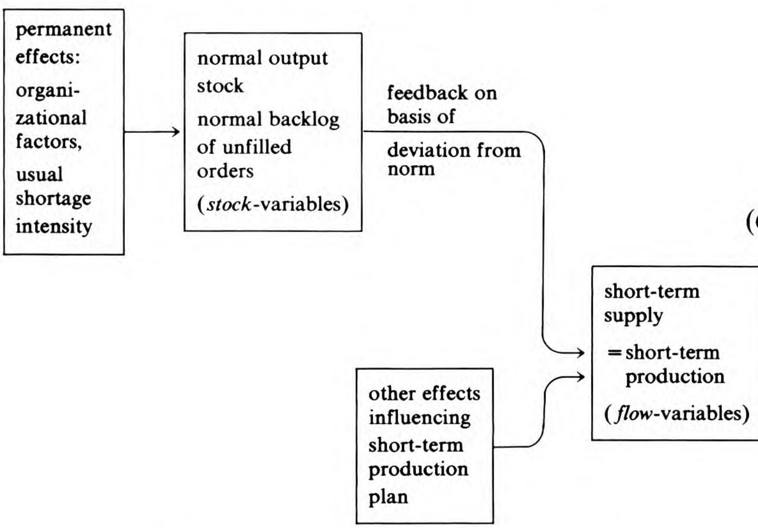
The first lesson is that we must beware of summary and rough judgements on stock reports. "Stocks have piled up – a bad sign." It may or may not be; it depends on which elements of total stocks have increased. If, for example, it is the output stocks of products the buyers like to purchase which increase, it indicates a *decreasing* intensity of shortage. If, however, it is input stocks of which more and more is hoarded, it is a sign of *increasing* intensity of shortage.

The second methodological lesson is that it is impossible to read off at a glance from aggregate statistics of stocks or of order backlogs whether we are dealing with a resource-constrained or a demand-constrained economy. The following kind of simple formula does not hold: a shortage economy can be recognized by its lack of stocks. Or, it is the distinguishing mark of a shortage economy that instead of (positive) output stocks, "negative output stocks", that is backlogs of orders appear in it. The fact that system A is resource-constrained and system B demand-constrained, must be revealed in the disaggregated *proportions* of input and output stocks, productive and unproductive slacks, and of order backlogs; more exactly, in the *deviation of proportions*. And this can be recognized only by comparisons which go into some detail.

6.5. Supply function

Now we can go on to discuss the short-term supply function. Here again, as in the case of the material demand function in Chapter 5, the causal chain determining supply will be divided into two elements. The first element is that persistent conditions determine the permanent characteristics of the behavior of the producer-seller. The second element is that these permanent features, and the circumstances at the given point of time together determine the instantaneous reaction of production.

In the case of supply the two elements are connected by the *norms* of the output stock and of the backlog of unfilled orders. Formation of norms and their characteristic magnitudes were discussed in the previous section. The chain of causal relations in represented in the scheme (6.1):



(6.1)

Relationship (6.1) outlines the characteristic supply response of a shortage economy. Efforts to increase production are concentrated where output stocks have decreased abnormally, or where the backlog of unfilled orders has grown unusually large, that is the queue is abnormally long. And, conversely, resources can be drawn away from wherever the output stock is abnormally large or the queue has become unusually short.

Of course, deviations from the normal output stock, or from the normal order backlog, together with demands from buyers do not represent the only impulses to which production and selling intentions react in the short run. In the *traditional* economic management system a decisive role is played by the plan directives received from superior authorities. Such directives are developed partly in response to similar impulses (stock and order signals), but they are also influenced by many other factors. In the *postreform* economic management system the firm becomes more responsive to relative prices of different outputs, as will be discussed in Chapter 14. These "other effects" are shown by the rectangle and arrow at the bottom of scheme (6.1).

The "function" (6.1) was here merely given verbally, without mathematical specification. (Later it will be specified for illustration, at least within the framework of a simple model.) The influence of the different explanatory variables remains an open question. In certain sectors only the output stock signal has any influence, while in other fields it is a signal related to the backlog of orders and in still other fields it may be both.

Our hypothesis is that *in the shortage economy the producer-seller is rather rigid. Short-term production and selling intentions usually react with lags and respond fairly weakly to the signals described in* (6.1). *Yet sooner or later, to a greater or lesser extent, a reaction does follow. If not especially sensitive to signals from the buyer, supply is not completely rigid, either.*

The hypothesis requires thorough empirical testing. The supply function (6.1) is not a mechanical rule of deterministic character. Supply does not "automatically" adjust itself to any signal arriving from the demand side. A supply reaction *may* follow, but it *does not have to* follow in every case.

Existing social systems, and within each system the various firms, may differ from each other in the speed and consistency of their reaction (6.1). *The more intensive the shortage, the slower the adjustment of supply to demand. The more the producer-seller enjoys a dominant position, the more he can leave it to demand to adjust to supply. And this then becomes one cause of continuing shortage. Exactly because supply adjusts itself badly to the buyer's requirements, products are often manufactured which the buyers do not particularly want and the goods they need most are not produced.* As in the previous chapter on the purchase side, we can now see on the sales side one of the self-generating vicious circles of shortage.

The supply function of the firm (6.1) has several properties which show a certain kind of symmetry to the demand function of the firm (5.4). A considerable part of what we said about demand functions in Chapter 5 could be repeated here, *mutatis mutandis*, about the supply function. We could also make some comparison with the neoclassical supply function, but we shall not do so because of limited space. However, in order to avoid misunderstandings, we must draw attention to two essential differences between the two relationships (5.4) and (6.1).

In the case of demand the *desired* level of input stocks was the intermediate term in the transmission, while in the case of supply it was the *norm* of output stocks (and of the backlog of unfilled orders). The desired level is, as we pointed out earlier, the decision-maker's aspiration level which may deviate considerably from the intertemporal average of the actual level.

The difference is not formal, but is based on experience. For the *seller* the normal level of output stocks or of the backlog of unfilled orders is more or less acceptable; he has no strong impulse to deviate from it. Therefore (if other factors do not influence him to do otherwise) he will adjust his short-term production plan to restore output stocks, or the order backlog towards the normal level. The situation is different for the *buyer* whose demand is almost-insatiable and who would like to accumulate the largest possible input stocks of those materials he needs most and whose purchase is the least secure. Therefore, he is always dissatisfied with the normal level of input stocks; he wishes to hoard more than that.[7]

A second difference is connected with the first. The function (5.4) explains the *instantaneous* buying intention, while function (6.1) shows the *short-term* production–selling intention. The difference is again explained by the fact that buyer and seller are in a different position. Chronic shortage keeps the buyer in a state of permanent vigilance: at every moment he has to reconsider what he should try to buy. Therefore, in describing demand we must go down to the instantaneous purchasing intention. On the other hand, the seller may rest assured that what he has in his store of output will be sold, too (except for a few totally "frozen" articles). His seller's role influences him in deciding on his short-term (yearly or quarterly) plans. Of course, he carries out instantaneous adjustment as well, but only in his producer's role (see Chapter 2). For this reason it is not important to go down to the instantaneous selling intention in explaining the chain of causal relationships.

[7]An important *methodological* lesson can also be drawn. Various kinds of control mechanism and various kinds of decision regularity exist: control according to *norm*, according to *aspiration level*, according to *critical value*, etc. Each can be described by its own model and own formalism. It is not necessary to demand a unified treatment of all these, but each behavioral pattern should be represented, as far as possible, by the model best reflecting it.

6.6. The seller's attitude

In the same way that we spoke of the buyer's attitude, we can also speak of a *seller's attitude*. What was said in section 4.7 to explain the word "attitude" also holds here. It would be unnatural to contrast all the components of the buyer's attitude to the corresponding components of the seller's attitude; there is no perfect parallel. There are, however, a number of components in which some contrast can be made. Let us recall some of the components of the buyer's attitude and contrast them with those of the seller's attitude.[8]

Demand function, including the desired size of input stock	Supply function, including the output stock norm, as well as the norm of the backlog of unfilled orders
The buyer's search propensity	The seller's search propensity
The buyer's waiting propensity	The seller's waiting propensity
The buyer's effort to win over the seller	The seller's effort to win over the buyer

Components of the seller's attitude together express the permanent features of the seller's behavior. Most of its components are observable and measurable.

Some of the components express the sacrifice one party is willing to make for the sake of completing the transaction. A form of sharing takes place, and the greater the sacrifice demanded of one party, the less is demanded of the other. It is a characteristic symptom of a shortage economy that the greater part of the sacrifice, burden, and inconvenience of a transaction is charged to the buyer. Only the most important will be mentioned.

(1) If search is involved, it is the buyer who visits all likely sellers. In the reverse case, that is in case of sales difficulty, it is the seller's agents who visit buyers.

(2) If it is necessary to wait, it is the buyer who waits for the product. In the reverse case, that is in case of sales difficulty, it is the seller who waits till the buyer happens to come to him.

(3) The buyer makes various kinds of efforts to win over the seller: he makes friends with him, rewards him with services, he may try to corrupt

[8] A few additional components of the attitude (e.g. the seller's sensitivity to output prices) will be discussed in Part II of the book.

him, he makes concessions as to the quality of the product, and so on. In the reverse case, that is in case of sales difficulty, it is the seller who "courts" the buyer, he tries to influence him by publicity, and win his benevolence by attentive service and extra thoughtfulness.[9]

The difference in attitude does not depend on the buyer's or the seller's mentality, good or bad manners. The sharing of burdens between them, the attitude towards each other, and the social relation between the two parties are determined by relative *power*. In a resource-constrained, suction economy the "seller's market" dominates, while in a demand-constrained, pressure economy there is a "buyer's market".

Relative market power has a strong impact not only on the relations between buyer and seller in the sphere of *trade* but also on *production*. What is more, over a longer historical period it is precisely the effect upon production that proves to be most important. In a "buyer's market" the producer-seller tries to win the customer not just by courtesy, but primarily by delivering faultless goods; and he tries to outrival his competitors by bringing out new products, better than old ones. That is the motivation that no longer operates when the firm sells its products without difficulty. In circumstances of chronic shortage the firm as buyer is prepared to accept even poor quality inputs. *The producer-seller feels no internal economic incentive to develop the market by introducing new products of better quality.*[10] *A shift in relative power to the seller's advantage strengthens impulses towards the quantitative increase of production while it weakens those favoring qualitative improvements in the product.*

Here we reach one of the gravest – or perhaps *the* gravest – disadvantageous consequences of chronic shortage. The drive which should enforce the continual qualitative improvement of production ceases to operate effectively.

[9]On the producer-seller's behavior and, in this connection, on the effect of the sales situation, see for example the article by Laki (1975).

[10]Here I only mention briefly this – perhaps – most important consequence of shortage, because the problem was discussed in detail in my book *Anti-Equilibrium* (Kornai 1971a, b; see, for example, the tables in ch. 20 on the introduction of new products).

CHAPTER 7

The normal state of production and trade

7.1. Introduction

After discussing the buyer's behavior in Chapters 4 and 5 and the seller's behavior in Chapter 6 it is now time to talk about the interaction between buyer and seller. First, in sections 7.2–7.5 we shall deal with a *partial* market;[1] that is with a single sector of trade in the customary framework of partial analysis. Then we shall go on to analyze the *general* interdependency of production and trade.

The present chapter differs somewhat in character from the preceding ones and, with the exception of Chapter 8, also from subsequent chapters of the book. We shall carry out the analysis on a *more abstract* level; *mathematical models* will serve as background for our propositions. Although we shall also try to represent real economic interrelations by these models and wish to promote clarification of problems raised in practice – we are compelled to make important compromises in the realistic description of practice. We shall use a number of abstractions and strong simplifying assumptions.

What we would like to obtain in exchange for such compromises is a "purer" form of theoretical generalization and a stricter deductive support for our propositions. We wish to demonstrate that the subject-matter of the present book, that is the examination of shortage economy and of non-price "quantity" adjustment, allows a number of questions to be studied with the aid of formalized models.

[1]The concept of market is interpreted in two ways. *In the broad sense*, it includes all transaction processes based on direct horizontal relations between supplier and recipient of the goods, even if price and money play little or no role in these processes. *In the narrow sense*, it is restricted to trading processes in which prices reacting to demand and supply play an important role. We use here the *broad* interpretation.

7.2. Partial analysis: Market for a product with a queue

As an introductory example we present some important interrelations in an extremely simplified model.[2] We describe the process of *instantaneous* adjustment.

Let us imagine a partial market, a narrow sphere of trade, where a single group of products is traded. This could be, for example, a certain machine part. Two concrete products belong to the group. Let them be two versions of the machine part in question: *good G* of better quality and *good H* of a poorer quality. Although they differ in quality, good H can serve as a substitute for G. There is usually a shortage of good G, requiring the buyer to queue for it. Good H is available at all times without queuing.

Service is arranged in such a way that the buyer gets one item of either G or H on each occasion, and after using it up, he must come forward to state his needs once again.

There is a single seller (e.g. the monopolistic manufacturer of the part in question). He faces a total of n buyers. (In our example they may be the purchasing agents of various factories.)

The model is dynamic: the buyers' demands for the part are constantly renewed; they come forward for it again and again; they buy it, use it, and the whole affair starts over again. This process will be described as a "closed circle": after purchase of the product the need to buy it again is renewed endogenously.

Let us recall Chapter 4 in which we accompanied a buyer on his shopping route and represented the series of decisions taken by him as an algorithm. We now take a similar course, but the actions of n buyers will be represented jointly, and their meetings with the seller will also be represented. Yet to simplify the model we omit for the time being a few of the decision points treated in Chapter 4. The process is demonstrated in fig. 7.1; we shall comment on each element in turn. The symbol $\otimes$ represents "switches", that is points where the buyer has to make a choice.

Let us begin at the left-hand side of the figure, with the lower symbol $\otimes$: that is the first decision problem of the buyer. Initial demand is formed. We assume that, considering quality differences (and perhaps also relative prices) between goods G and H some buyers, more exactly a fraction λ of them, decide to try to buy good G. The remainder, that is a fraction $(1-\lambda)$

[2]This model is presented – somewhat extended and with less restrictive assumptions – in Mathematical Appendix A, whose co-author is Jörgen W. Weibull. In sections 7.2–7.5 I have made extensive use of results from the research work carried out jointly with him.

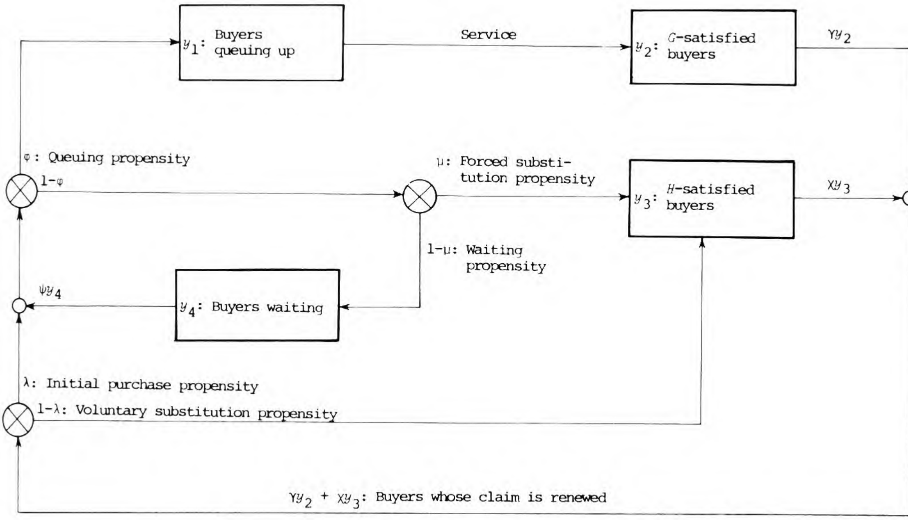


Figure 7.1. Dynamic model of a partial market.

carry out voluntary substitution right at the beginning, substituting H and G. It is assumed that their decision was not due to purchasing difficulties or to the length of the queue. The ratio of the choice between the two alternatives would be the same even if both goods G and H were available without supply constraints or queuing. The parameter λ is the *initial purchase propensity* of the buyers, and $(1-\lambda)$ is the *voluntary substitution propensity*.

Let us now follow the arrow pointing upwards from the left-hand corner. Those trying to buy good G have arrived at the seller's and catch sight of the queue. Let us denote *queuing time* by W. There is another symbol $\otimes$: a new decision problem. The question arises for the buyers of whether they should join the queue and await their turn. Let us denote the buyers' *queuing propensity* by $\varphi(W)$, a decreasing function of queuing time. The longer one has to wait for service, the smaller the number of buyers willing to join the queue. With a given queuing time W a fraction $\varphi(W)$ of the group trying to get good G will join the queue, and a fraction $(1-\varphi(W))$ will not.

The *number of those queuing up* at time t is denoted by $y_1(t)$. They will wait out the period W and get the desired product in the end. During a given unit of time the seller can serve at most s buyers. This is called the seller's *supply rate*.

With the freshly acquired product in hand, the buyer goes away and feels satisfied for a while, until his need is renewed again (for example when the part has worn out). The time period between purchase and revived demand for good G is denoted by $1/\gamma$. Buyers in this state at time t are called *G-satisfied* buyers and their number is denoted by $y_2(t)$. From this group of buyers a fraction γ comes forward with a renewed need for the good. They appear at the right-hand side of the figure.

And now let us return to the buyers who were not prepared to join the queue immediately, because they were deterred by the lengthy waiting time. They are faced with a decision dilemma, as shown by the symbol $\otimes$. Some of them bring themselves to buy good H. What they did not do at first voluntarily, they now do have to do, as an effect of shortage. Let us denote by μ the buyers' *forced substitution propensity*. This means that a fraction μ of the group of buyers that set out to purchase G, and find the given queuing time W unacceptable is prepared to accept forced substitution. This is clearly distinguished from the voluntary substitution propensity $(1-\lambda)$ which relates, by the way, to a larger group: all those whose need for a new good (G or H) is renewed.

Those who have undergone forced substitution, together with those who chose good H voluntarily (they appear at the center of the figure), must consider themselves satisfied for a while. Many of them may feel less fortunate than if they had secured good G. Yet they can manage somehow with H. The time period between purchase and the renewal of need for good H is denoted by $1/\chi$. (The product being inferior, this may well be shorter than the G-satisfaction period.) Buyers who are in this state at time t are called *H-satisfied* buyers and their number is denoted by $y_3(t)$. From this group of buyers a fraction χ comes forward with additional demand for good H during a given unit of time. These will join their companions, who used good G, at the switch to be seen at the right-hand side of the figure.

There are also, however, some buyers who, although they were not prepared to join the queue immediately, are not prepared to accept forced substitution either. They choose rather to go home, wait passively, and after a time visit the seller again and have another look at the length of the queue. Let us denote by σ the *waiting propensity* of the buyers. (This is passive waiting: postponement of the dilemma "to join or not to join the queue".) A fraction σ of the group of buyers unwilling to join the queue immediately chooses to postpone a decision. The number of *passively waiting* buyers will be denoted by $y_4(t)$. From this group of buyers, during a given unit of time, a fraction ψ pays another visit to the seller to inspect the length of the queue and form some idea about the likely queuing time. In other words the *waiting time* is $1/\psi$. This having passed, the buyer returns to the decision dilemma $\otimes$ in the center of the figure.

In the present simple form of the model we assume that there are no other alternatives for choice. It follows that those not prepared to join the queue immediately are either passively waiting, or perform forced substitution: $\sigma = 1 - \mu$.

Thus, we have described the full cycle of the market process.[3] As was mentioned earlier, the complete sequence of purchase, utilization, renewal of need, and decisions preceding actions has been presented as a *closed circle*. The mass of buyers is permanently flowing in this circle. Every buyer is in one of the following four kinds of state at each point of time:

$$y_1(t) \quad + \quad y_2(t) \quad + \quad y_3(t) \quad + \quad y_4(t) \quad = \quad n. \tag{7.1}$$

number of buyers queuing up	number of G-satisfied buyers	number of H-satisfied buyers	number of waiting buyers	total number of buyers

Summing up, we can state the following: our model describes *instantaneous (physical) supply* by the supply rate s on the one hand, and it represents the *buyers' attitude* on the other. It is true that this is done in a very simple form, yet in the spirit outlined in Chapter 4. Components of this attitude are listed below:

$\lambda, \gamma, \chi =$ generating *initial demand*;

$\varphi(W) =$ *queuing propensity*;

$\mu =$ *forced substitution propensity*;

$(1-\mu), 1/\psi =$ *waiting propensity* and *waiting time*.

The components mentioned above (one function and some constant parameters) express the *average* attitude of the total number n of buyers. Thus, for example $\mu = 0.6$ means that out of 100 buyers that wished to buy G but were not willing to join the queue with the given queuing time, an average of 60 are willing to accept forced substitution and 40 choose to wait and try again later.

[3]We dispense with a few further assumptions whose detailed description is given in Mathematical Appendix A.

7.3. Normal state of the market: First approximation

It can be proved that *a steady state exists* on the partial market described in the previous section.[4] And, since certain characteristics of the market are constant over time (the number of sellers and buyers is unchanged, and so is the supply rate), this is a *stationary* state. If trading activity is in this state, and no external circumstances disturb it, it will – from the effect of its intrinsic regularities – perpetually repeat itself. The queue will always be as long, and waiting in the queue will always last the same time. The number of buyers wishing to purchase good G, of those carrying out voluntary and forced substitution, those passively waiting, and of those in the course of using goods G and H, will all be constant. This is called the *normal state* of the particular market which we described in the preceding section.

Just as earlier, the adjective "normal" does not imply a value judgement. For example, the normal state may involve a normal queuing time of $W^* = 3$ hours. Those who spend their time in the queue certainly do not enjoy it. What we assert is solely that the state in which the queuing time is three hours can be permanently reproduced.

It is a market not in *Walrasian equilibrium*: there is a chronic shortage of good G. Initial demand for good G cannot be satisfied for all buyers; queuing and forced substitution are permanent. It may even be *very far* from Walrasian equilibrium; the queue may be very long, and the extent of forced substitution considerable. Yet this is a normal state, since it reproduces and perpetuates itself and because its participants themselves accept it as a normal state.

This brings us to one of the key questions of our discussion. A normal state of the market, far from Walrasian equilibrium, can be sustained by several different factors. One reason is undoubtedly that participants acknowledge the existence of this normal state. They grumble sometimes, but in the end resign themselves to it and adjust their attitudes accordingly. A fraction $\varphi(W)$ of the buyers wishing to buy good G is *willing* to join the queue; of those who do not enter the queue, a fraction μ is *willing* to accept forced substitution, and so on. *As long as the attitude of participants in the market process is unchanged, the normal state can be perpetually reproduced. And, if that attitude has adjusted to conditions of chronic shortage, the chronic shortage can consolidate.* This proposition could only be proved

[4] The article by McCafferty (1977) talks about the "excess demand steady state" in a similar sense.

strictly for a few specific models so far; later we shall revert to it, with less restrictive assumptions.

Of course, various factors may cause the *instantaneous* state of the market to deviate from the normal state. It can be proved that the partial market outlined in the previous section has a kind of *stability*. Deviating from its normal state, it will sooner or later return to it. For this a *control mechanism* is needed, with a *signal* which drives the system back towards its normal state. In our case the queuing time W serves as such a signal. If the actual queuing time is longer than normal, it will discourage further buyers from queuing; more of them will choose forced substitution or waiting and postponing the decision. This also modifies the dynamics of the renewal of needs. As a result the queue will grow shorter and queuing time return to normal. And, if queuing time is shorter than normal, the reverse mechanism will take effect: more buyers will be encouraged to join the queue, and so on.

Although here, in this very simple model taken as an introductory example, only a single "feedback" is functioning, namely the orientation of buyers by waiting time in the queue, a general conclusion can be drawn. *A normal state is not simply the intertemporal average of certain attributes of a process. The term "normal state" of a system is justified only if a control mechanism exists which – in the case of deviations – repeatedly drives the system back to this normal state.* The market in a shortage economy is a system exactly like that: control mechanisms operate in it, inducing it to reproduce repeatedly the customary degree of shortage.

At this point I wish to make a short detour in the history of economic thought. Already in Chapter 2 we quoted the Soviet disputes of the 1920s, in which the "general overproduction" of the capitalist economy was contrasted with the "general shortage" of the socialist economy. The authors cited there felt that the shortage intensity reached such a high degree that it resulted in a special kind of *crisis*. "...in the commodity capitalist society the crisis appears in the form of an *overproduction crisis*, while in the natural proletarian society it appears in a contrary form: in the form of an *underproduction crisis*", wrote Kritsman.[5] And Novozhilov said, "...instead of a general overproduction crisis a general commodity shortage crisis develops...".[6]

The economists of the 1920s experienced – on a historical scale – the first shock of "general shortage". The approach of the present book differs

[5]See Kritsman (1925, 1929).
[6]See Novozhilov (1926).

essentially from this. *Shortage is not a "crisis phenomenon"*. Chronic shortage is the *normal* state of the resource-constrained economy. It is compatible with its normal operation and growth; indeed, it is not just "compatible" with it, but it is one of the permanent features of its normal operation.[7]

7.4. Shifts in the normal state

A normal state does not last forever. It may shift if an essential and permanent change takes place in social conditions or in the other circumstances determining the state of the market. A *long-term* shift or displacement of the normal state must be distinguished from the previously discussed *instantaneous* fluctuations around the normal state of the market. Let us assume that on our partial market the normal share of forced substitution is 50 percent. The actual share may be zero one day, and 100 percent the next, yet the average share over a longer period would be 50 percent (and most days it varies between 40 and 60 percent). This instantaneous fluctuation is to be distinguished from a long-term change in the share; if, for example, the normal share of forced substitution was reduced to 20 percent.

Obviously, the decisive factor here is the *long-term relationship between demand and supply*. Let us recall the example of section 7.2. Buyers were regularly queuing up for good G. The queue could be entirely eliminated in two ways. In the first, the supply reaction discussed in Chapters 3 and 6 takes place. The producer reacts to the signal that a queue always forms for G but never for H, and changes his output composition, producing more of G and less of H. The second possible change is that users begin to dismantle the machines for which they have been buying spare parts G. Demand diminishes and the queue shrinks.

The shift in the normal state of one or other partial market may be related to a *long-term change in the sellers' and buyers' attitude*. There was a period in Hungary when an electricity shortage was considered more or less customary, while labor was flowing abundantly into factories. Today an electricity shortage would cause surprise and indignation, while the labor shortage has grown chronic. Getting accustomed to one or other

[7]The instantaneous intensity of shortage fluctuates around its normal value, and from time to time may grow particularly severe, which may lead to the sharpening of social and economic tensions.

manifestation of shortage does not mean that it does not cause annoyance and difficulties, only that it is part of the habitual order of things. It is difficult to say whether the change in attitude is a cause or a consequence of the change in the normal state, since we have here a two-way cause–consequence interaction. In most cases the change in market conditions forming the attitude is the primary phenomenon and the adjustment of the attitude is secondary. As already mentioned, it is precisely the persistence of this attitude, that is the conformity of market participants, that promotes consolidation and self-perpetuation of the normal state. Yet sometimes the direction of the cause–effect chain changes. It may happen that participants of the market process are so extremely dissatisfied that they protest against the shortage phenomena and try to put pressure on the decision-makers to change the situation.

It may be that these changes take place unevenly in different fields of the economy. While on one market the buyers' forced substitution, queuing, and waiting propensity noticeably diminishes, on other markets the buyers may be more tolerant. In such instances, perhaps, the normal intensity of shortage decreases on the first market, while it increases on the other. This question would, however, lead us on to the problem of intersectoral linkages, as well as to the description of economy-wide shortage situation; these will be treated in the second half of this chapter and in later chapters of the book. For the time being let us remain with partial analysis.

7.5. Extension of propositions concerning partial markets

In the foregoing we discussed a number of propositions in the framework of a partial analysis, to do with (1) the existence and stability of the normal state of a partial market in a shortage economy; (2) mechanisms driving the instantaneous actual state of the market towards the normal state; and (3) long-term shifts in the normal state. All these propositions could only be proved deductively so far for the simple model described in section 7.2. We conjecture that similar assertions hold in a much wider sphere. They are certainly valid for a partial market in a shortage economy, within the framework of a model with less restrictive assumptions.

(1) The propositions seem to be valid not only for a *stagnating*, but also for a *growing* economy. For this proposition, characteristics of the normal state need to be reformulated. Indicators must be specified which are independent of the trade volume of the sector in question, and indepen-

dent of the number of sellers and buyers, and so on. The buyers' situation can be described, for example, in the following way: the ratio of the volume of products purchased by forced substitution to the total volume of purchases; the ratio of the number of buyers queuing to the total number of buyers, and so on.

The data mentioned above are percentage indicators. Let us call them *relative* indicators,[8] as opposed to the *absolute* indicators of the volume of forced substitution or the number of those queuing. Relative indicators may serve as a basis for comparison of the same market at two different points of time, even if the absolute attributes have changed between the two dates (e.g. volume of trade has grown). Similarly, two different partial markets can be compared at the same date, without regard to differences in absolute indicators.

Now a more general definition can be given. *The market is in a normal state if its characteristic relative indicators are unchanged over time.* In a special case (stagnating market, stationary process) this may also coincide with constancy of the corresponding absolute indicators.

(2) In section 7.2 buyers' attitude was considered to be homogeneous and only their average attitude was described. Instead, an assumption closer to reality can be used. The population of buyers can be partitioned into subgroups, and it can be assumed that the attitude of buyers is uniform within any given subgroup, while differing between outgroups. For example, the first type of buyer may be less patient than the second type: he may be willing to accept more forced substitutions, but does not wish to wait passively. Accordingly, $\mu_1 > \mu_2$ and $\sigma_1 < \sigma_2$. Then it can be proved theoretically that if the distribution of buyers by type of attitude is unchanged, the above propositions about the normal state will remain valid.

(3) Our introductory example presented a deterministic model. In reality it would be more correct to describe buyers' and sellers' attitude by random variables. These random variables have stable probability distributions, representing unchanging attitudes.

[8]Attention is drawn to the fact that neither in reality, nor even in the theoretical model do percentage rates expressing buyers' *propensities* coincide with the corresponding percentage rates expressing the *state occurring* on the market. Reverting to an earlier example, let the forced substitution propensity be $\mu = 0.6$. The ratio of buyers actually effecting forced substitution to the total number of buyers, i.e. y_3/n, may still be either larger or smaller than 0.6. This depend on numerous factors, including the number of buyers joining the queue. In our model "propensities" are the parameters of the equation system, while "state variables" are the unknowns in the same equation system. It is true that the latter depend on the former, but the dependence asserts itself through complicated transformations.

In the case of a stochastic representation the state of the market is also described by random variables. In this framework *the normal state of the market can be described by the distributions (and especially by the expected values) of the main state variables.* Thus, for example, the queuing time, $W(t)$ is a random variable for each t. Its normal value, W^*, is the expected value of this random variable.

(4) In our introductory example a single seller worked in the market. The structure of real markets is often different, with several, or even a large number of sellers present in the market. Accordingly, as discussed in Chapter 4, one alternative for the buyer in the case of shortage is *search.* He goes from one seller to the next trying to obtain the product. One component of his attitude is his search propensity. And now the state variables of the market are completed by actual search (e.g. searching time, or the number of buyers searching).

(5) In our introductory example the supply rate is unchanged over time. It is exclusively the buyer who adjusts himself to the market situation, while the sellers' behavior is rigid. In reality supply may, of course, also be controlled by a feedback mechanism, as was already mentioned in Chapters 3 and 6.

(6) In the deterministic model of the introductory example there is no slack in the normal state with a permanent supply rate. While there is a queue, the seller can always sell his entire supply. Yet in reality goods do not come into stock regularly, and for short periods the queue may even disappear. In Chapter 6 we have already treated time lags between the arrival of goods and their transfer to the buyer, which lead to the formation of a transitory product stock. In reality this already provides sufficient reason for the phenomenon mentioned several times: considering a long period as a whole, *slack and shortage may exist simultaneously.* Under such circumstances, an important characteristic of the state of the market is the size of the slack. For this it is also useful to employ relative indicators: the proportion of the slack to total trade. And, if a market for a large group of products is described, it is particularly important not to use only the aggregate slack indicators but also some suitably disaggregated ones. It was mentioned in the preceding chapter that, while "frozen" stocks of certain products may accumulate, all the stocks of other products may be sold out in a short time as a consequence of shortage. It would not be correct to make these deviations disappear through "averaging".

A few steps have been taken in the direction of strictly proving our assertions about the existence of a normal state of the market, its stability, control mechanisms, and so on, with the aid of theoretical models less restrictive than the simple example presented in section 7.2. Preliminary results have been achieved mainly in the topics mentioned in paragraphs (2), (3), and (4): in the partition of buyers into subpopulations, the stochastic description, and the consideration of search.[9] Theoretical research must continue in the direction of developing models with fewer restrictions, coming closer to reality.

Of course, our conjectures not only require abstract theoretical verification but, above all, *empirical testing*. Sporadic observations and partial data seem to support our hypotheses. It appears that the partial markets of a shortage economy have in fact a characteristic long-term normal state: the customary ratio of forced substitution, the typical queuing time, and so on. As mentioned several times, control mechanisms that repeatedly drive partial markets back to their normal state, if they have deviated from it, can be observed in practice. The task of empirical testing leads us to problems of *measurement*, which will be discussed at the end of this chapter.

7.6. General interdependence: Leontief economy

After analyzing the operation of a partial market in sections 7.2–7.5, we now turn our attention to interdependencies between partial markets. We examine a system in which there are m products, each of which can be used as an input for other products. Real processes are directed by such nonprice control mechanisms as have been treated in earlier chapters. Does the system have a normal state? The question to which an answer is sought is in a certain sense related to the sphere of problems investigated by general equilibrium theory. In this chapter, too, we analyze the "general" normal state – in the sense that we examine the general interdependence between the individual producers and consumers of products.

The elaboration of a comprehensive theory of multiactor, multiproduct, interdependent, abstract economic systems with nonprice "quantity" adjustment is only at its earliest stages. The present book indicates, in outline, only a fraction of the theoretical results already achieved.[10] We shall be content just to give an idea of the approach to the problem.

[9]Some of the results are discussed in Mathematical Appendix A.

[10]In sections 7.6–7.8 I relied largely on the research begun together with Béla Martos and continued with the participation of András Simonovits and Zsuzsa Kapitány. (See Kornai–Martos, 1971, 1973; Kornai–Simonovits, 1975a, 1975b, 1977a.) Results of the research work are summarized in the collection of papers by Kornai–Martos (1979a, b). At this point we only give the outlines of a few ideas which are explained in detail in the collection.

We start from an extremely simplified model: a Leontief economy.[11] Each product can be manufactured only by a single monopolistic firm, and by only one possible technology. A rigid complementarity asserts itself between inputs. There is no substitution between products, neither voluntary nor forced. I am aware that in this way a considerable range of shortage problems has been excluded, but we must apply this drastic simplification in order to proceed.

The system is dynamic; time is a continuous variable. We examine the instantaneous adjustment processes of production and the use of inputs. In the description that follows the time-argument of each variable will not be indicated.

The real sphere, that is the physical input–output relations in the system, are described by two balance equations. One is the *balance equation for the input stock*:[12]

$$\underset{\boxed{\text{change in input stock}}}{\dot{v}_{ij}} = \underset{\boxed{\text{purchase}}}{y_{ij}} - \underset{\boxed{\text{input coefficient}}}{a_{ij}} \quad \underset{\boxed{\text{production}}}{x_j}, \qquad i=1,\dots,m; \quad j=1,\dots,m. \tag{7.2}$$

The other is the *balance equation for the output stock*:

$$\underset{\boxed{\text{change in output stock}}}{\dot{u}_i} = \underset{\boxed{\text{production}}}{x_i} - \underset{\boxed{\text{sales to firms}}}{\sum_{j=1}^{m} y_{ij}} - \underset{\boxed{\text{sales to households}}}{g_i}, \qquad i=1,\dots,m. \tag{7.3}$$

Although Part I does not deal, as a rule, with the household sector, in the present model – representing the general interdependencies of all economic sectors – the household sector does appear, at least in an aggregate form. Household demand is exogenously given.

As I stressed in Chapter 1, the above-mentioned collection of papers and the present book are "joint products". The collection is in numerous respects *complementary* to the present book.

Background for the model discussed in section 7.6 is Kornai–Martos (1971, 1973); the rest of the chapter also makes use of the other works mentioned.

[11]From the literature treating the Leontief model we must mention Leontief (1953, 1966, 1977), and Bródy (1964, 1969, 1970).

[12]The dot over the symbol for a variable indicates the derivative with respect to time.

The control sphere, that is the control mechanisms operating in the system, are also described by two functions. One is the *shopping rule*, which could also be called the *demand function*:

$$\underbrace{\dot{y}_{ij}}_{\text{change in purchase}} = \underbrace{a_{ij}\dot{x}_j}_{\text{change in producer utilization}} - \underbrace{2\Theta\Lambda}_{\text{control parameters}} \underbrace{\dot{v}_{ij}}_{\text{change in input stock}} + \underbrace{\Lambda^2}_{\text{control parameter}} \Big(\underbrace{v_{ij}^0}_{\text{desired input stock}} - \underbrace{v_i}_{\text{actual input stock}}\Big), \quad i=1,\ldots,m; \quad j=1,\ldots,m. \tag{7.4}$$

According to the rule, the producer firm will increase its purchase if (1) its production has grown, (2) its input stock has diminished, and (3) its actual input stock has fallen below the desired level.[13] These are simple and sensible behavioral rules, revealed in every enterprise manager's attitude in every economic system. And, if price signals have no or hardly any effects, the stock signal assumes a particularly important role.

Even though the chosen mathematical form is quite specific, the general contents of the signal–reaction relationship are consistent with what has been explained in the book so far (particularly in Chapter 5) about the behavior of the firm as buyer, the formation of its demand, the properties of its demand function, and of its purchase.

The other function is the *production rule*, which could also be called the *supply function*. It will be presented in two versions: first in the form (7.5), according to which production reacts to stock signals; and later it will be given in a modified form, in which the backlog of unfilled orders serves as

[13]The above-mentioned economic interpretation of the exogenous magnitude, v_{ij}^0, appearing in the purchasing rule is consistent with what we said about the demand of the firm in Chapter 5 – but it differs from the interpretation of the same magnitude as given in the study by Kornai–Martos (1971, 1973).

a signal:

$$\underbrace{\dot{x}_i}_{\boxed{\text{change in production}}} = \underbrace{\sum_{j=1}^{m} \dot{y}_{ij}}_{\boxed{\text{change in sales to producer firms}}} + \underbrace{\dot{g}_i}_{\boxed{\text{change in sales to households}}} - \underbrace{2\Theta\Lambda}_{\boxed{\text{control parameters}}} \underbrace{\dot{u}_i}_{\boxed{\text{change in output stock}}}$$

$$+ \underbrace{\Lambda^2}_{\boxed{\text{control parameter}}} (\underbrace{u_i^*}_{\boxed{\text{normal output stock}}} - \underbrace{u_i}_{\boxed{\text{actual output stock}}}), \quad i=1,\ldots,m. \tag{7.5}$$

According to rule (7.5) the firm will increase its production if (1) its sales to other producer firms and households have grown, (2) output stock has diminished, and (3) actual output stock has fallen below its normal level. As said previously in connection with purchases, these are simple and sensible behavioral rules which hold generally, independently of the specific social system. And we must add once more that if price signals have little or no influence, the effect of stock signals may be quite strong.

In its general content control by norms according to formula (7.5) is consistent with what has been said in this book so far (particularly in Chapters 3 and 6) about the behavior of the firm as seller, the formation of production and supply, the properties of its supply function, and of its sales.

It can be proved theoretically[14] that *the system* described in eqs. (7.2)–(7.5) *is viable. It is capable of continually reproducing itself, while satisfying household demands. A normal state exists in which the desired and actual values of input stocks, and the normal and actual values of output stock variables, coincide. It has a certain degree of stability: if it moves away from the normal state, its control mechanism is able to drive it back to normal.*

The proposition formulated above is also valid if it is not positive output stocks but "negative output stocks", that is the backlog of unfilled orders

[14]Other complementary assumptions necessary for proving the propositions are given in Kornai–Martos (1971, 1973).

that appears in the feedback, in other words, if the *stock-signal* production rule (7.5) is replaced by an *order-signal* rule.

For the sake of brevity, let the household sector be buyer number $(m+1)$. *The balance equation for the backlog of unfilled orders* is then the following:

$$\underset{\boxed{\text{change in the backlog of unfilled orders}}}{k_{ij}} = \underset{\boxed{\text{order}}}{l_{ij}} - \underset{\boxed{\text{sales}}}{y_{ij}}, \qquad i=1,\ldots,m; \quad j=1,\ldots,m,m+1. \tag{7.6}$$

Summing over all buyers:

$$k_i = \sum_{j=1}^{m+1} k_{ij}, \qquad i=1,\ldots,m. \tag{7.7}$$

The feedback signal in the production rule of firm i is simply:[15]

$$(\underset{\boxed{\text{normal backlog of unfilled orders}}}{k_i^*} - \underset{\boxed{\text{actual backlog of unfilled orders}}}{k_i}). \tag{7.8}$$

Although theoretical proofs have been developed for the two kinds of "pure" system: *either* the stock-signal mechanism *or* the order-signal one, we conjecture that their validity can also be extended to cover "mixed" cases, in which both kinds of signal are operating.

[15]Kornai–Martos (1971, 1973) prove the viability of the system as well as its other previously discussed properties for the case of the stock-signal production rule (7.5). Kornai–Simonovits (1975b) describe similar properties for a related although somewhat different system, with purchase and production rules specified slightly differently. We do not wish to trouble the reader here with a description of the differences.

7.7. Permanently away from Walrasian equilibrium

Now we have a system before us, even if only as a sketchy description, which resembles a real shortage economy in two fundamentally important features: *(1) buyers have to wait before their orders can be met, and (2) production and purchases react exclusively to nonprice "quantity" signals (stock- and order-signals). Producers and consumers of various products may be coordinated with each other even in such a system.* The normal intensity of shortage may consolidate, represented by the normal backlog of unfilled orders. And normal slack may also consolidate: large input stocks as a result of the hoarding tendency, and low output stocks as a result of almost-insatiable demand.

A normal state of production and trade might be generated, which is permanently away from Walrasian equilibrium, because shortage and slack have consolidated in it.

I would not like to mislead the reader concerning the theoretical maturity of the above propositions. Chapters 2–6 of the book tried to characterize in detail the producer firm's behavior as well as that of the buyer and seller firms separately, each basically in verbal form. In sections 7.2–7.5, analyzing the interaction of participants in a partial market model, the results were based to a large extent on mathematically established propositions (complemented by a few more general conjectures). This was done, however, on the basis of models which were much "poorer" than the verbal description of Chapters 2–6, with considerable simplification in comparison to that. And now, in section 7.6, the sphere of analysis has become even wider. Turning our attention to the interdependencies among *m* products and still relying on mathematically established results, we had to retreat further as far as the "realism" of the model was concerned. (Leontief economy, exogenous household demand, etc.) I believe that the model-family outlined here could be developed further and that our propositions could be proved under much more general conditions. This will be a task for future research.

Among several weaknesses of the model developed in section 7.6, one of the most serious is the assumed rigidity of input combinations. Yet in earlier chapters we emphasized that one key to an understanding of shortage is the study of forced substitution. Here we can only express the hope that, having succeeded in proving the possibility of a continuous self-reproduction of "normal forced substitution" in the framework of a *partial* analysis, we shall also succeed, sooner or later, in proving the same thing for an interdependent multiactor, multiproduct system. Behavioral rules, such as were treated in sections 3.3, 4.2, and 7.2 in connection with forced substitution, should be built into multiproduct models as well.

At this point in our train of thought, it seems appropriate to make – as a detour – a few remarks on the concept of "equilibrium".

The *broad* interpretation of equilibrium – in general use in the natural sciences – is the following: it is that state of a system to which it always returns on account of its own regularities. The system is in equilibrium if the forces operating in it mutually counteract and compensate each other. Here equilibrium is a descriptive category. It leaves entirely open the question of whether this equilibrium – meant in the broad sense – of the system in question is to be considered "good" or "bad" by anybody (either an external observer or a participant in the system). The forest as a system of living things is in equilibrium if the wolf devours the hare; if it did not do so, the predatory animal would perish and so the customary proportions between the various species of animal would be disturbed. This event is part of the equilibrium even though the hare obviously finds it "bad". Counterforces operate to offset the actions of the predator and ensure the normal proportion of hares in the animal population. Such forces include the defensive and self-preservation activities of the hares (e.g. they try to escape from the wolf), and their appropriate rate of multiplication. A synonym for equilibrium in its broad sense – and an exact and equivalent synonym – is the expression "normal state".[16]

The category of equilibrium is applicable to social systems in a similarly wide sense. Its precise conditions are always system-specific. They depend on the state which is permanently reproduced by the general economic laws and by the system-specific intrinsic regularities.

Equilibrium in the broad sense (that is to say, a normal state) is also a *descriptive* category in social science. We neither praise nor condemn a system by saying that it is in equilibrium, that is "normal". In addition, this statement does not imply that the members of the system in question, or even the majority of them, are satisfied with the system. A social system may be in equilibrium even if a large number of its citizens are dissatisfied – but they are dissatisfied for different reasons, and so their actions have opposing effects. An equilibrium establishes itself if these

[16]The term, "normal state" has spread in natural science and in the philosophy of general systems theory inspired by the natural sciences. (See, for example, Koehler, 1938.) On the interpretation of the notion "normal state" and its tradition in economics going back to Marshall, see Shackle (1972).

opposing forces counteract each other; even though some internal tension persists, an enduring compromise develops in actual conduct.

Therefore, if we try to clarify in regard to some social system, whether it is capable of being in equilibrium, we always have to analyze whether there are internal control mechanisms or social regularities which – as trends or long-term tendencies of deviations and fluctuations – bring about and permanently reproduce the normal state. It is not by relying upon criteria brought into the analysis from outside by an observer but the specific internal regularities and norms of the system that allow us to characterize the normal state of a particular system. It is part of the normal state of an Asian society stagnating for hundreds of years that it is stagnating. It is part of the normal state of a shortage economy that there is chronic shortage in it.

Real social systems do not have a single specific normal state, but rather a set of normal states may be imagined. At this point we do not wish to define exactly the concept of normal state beyond what has been said, or to clarify the relation of this concept to similar concepts from the mathematical theory of dynamic systems (e.g. to the various types of stability). Here we only intend to make the concept intuitively clear. Thus the category, "normal state", is a collective term, which has to be specified in the course of each particular application, or be defined precisely in the framework of some mathematical model.

A system is in a *nonequilibrium* state if transitorily, as an accidental fluctuation, it deviates from its own intertemporal trend, its own norms, that is its own equilibrium understood in the broad sense.

This broad interpretation of equilibrium is not unknown in economics. For example, we commonly use the expression, "Keynesian unemployment equilibrium". The state of a capitalist economy analyzed by Keynes, in which mass unemployment becomes chronic, is thereby understood.[17]

In economic theory, however, as well as in everyday economic life, a much *narrower* interpretation of the concept of equilibrium is widely adopted. According to this the economy – and within the economy mainly the market – is in equilibrium if supply meets demand. Many call this a *Walrasian* equilibrium, although the terminology is not to be considered generally accepted.

[17]The category of equilibrium is used in this broad sense in the studies by Benassy (1974, 1975) and Malinvaud (1972), analyzing the persistent states of various systems deviating from Walrasian equilibrium (permanent unemployment, permanently repressed inflation, etc.).

Let us remain first at the level of pure theory, that is of models. Walrasian equilibrium is then a *special case* of equilibrium in the broad sense. It is a well-defined special case; we are familiar with the models in which the equilibrium state is a Walrasian equilibrium. There are, however, other theoretical models in which *non-Walrasian* equilibrium appears. Thus, for example, in the world of the models treated in the present chapter, the normal state is conspicuously non-Walrasian. In our partial market model queuing and forced substitution are features of the normal state; in our multisectoral model the backlog of unfilled orders may become a permanent characteristic.

If we leave the world of models to enter into reality, we can recognize that there does not exist an economic system with a Walrasian equilibrium as its normal state. Every system – not only in its instantaneous fluctuations, but in its intertemporal average as well – departs from the strict Walrasian equilibrium. This point will be treated in more detail in Chapter 8, and in subsequent parts of the book. It is, however, highly characteristic of any existing system, how, in which direction, and how far its normal state deviates from the Walrasian equilibrium.

I think that the terminology just introduced is clear and unambiguous. Yet it is not to be expected that everyone will accept it. Terminological confusion seems almost inevitable, because usages different in two respects from those proposed above are deeply rooted in the minds of most theoretical and, indeed, practical economists.

First, many identify the *special*, narrow concept (Walrasian equilibrium) with the *general*, broad concept (equilibrium in the natural scientific sense, system-specific equilibrium, normal state).

Secondly, many attach a value judgement to the word "equilibrium". If a system is not in equilibrium it is "bad". This is, of course, related to the problem first mentioned. It is considered obvious that equilibrium means Walrasian equilibrium, and any state in which supply does not satisfy demand is disapproved of.

I would not like to be unduly dogmatic about this terminology. I shall be satisfied if the readers of this book can follow clearly what I have to say, without terminological confusion. To facilitate this, the book has been using, and will continue to employ, the following terminology.

I shall not use the expression "equilibrium" in the broad sense (though I consider it theoretically correct); instead, I shall always write the words, "normal state".

I shall, however, use the expression "Walrasian equilibrium" in its *narrow*, special sense and the attribute "Walrasian" will always be present.

I have all the more reason to be cautious, because I must acknowledge that my book, *Anti-Equilibrium*, was itself insufficiently precise in its interpretation of the word "equilibrium". On the one hand, what I said at the beginning of the present discussion, namely that there exists an equilibrium in the wide sense, was not clearly stated. The reader could have formed the impression that I was denying it. And yet to deny it would be almost equivalent to rejecting the idea that there exist in each system deeply rooted intrinsic regularities which constantly reproduce the essential properties of the system.

On the other hand, I personally more or less fell into the terminological error of confusing the broad and narrow, general and special interpretations of equilibrium. Thus, for example, I referred to pressure and suction as states of permanent *dis*equilibrium, which suggest that the "genuine" equilibrium is the Walrasian one.[18] Now, as is clear from the preceding explanation and the whole spirit of the present book, I consider pressure as well as suction to be normal states of an economy, under suitable conditions.

7.8. Vegetative control

And now, after our terminological digression, let us return to the models outlined in section 7.6 and examine their structure of decision-making and information-flow.

In the purely stock-signal economy control is decentralized in two different respects. On the one hand, every firm makes its *decision* regarding both production and purchases independently, without agreement with other firms. On the other hand, in its decisions it uses exclusively decentralized *information*. Let us consider once again the rules (7.4) and (7.5). The firm reacts only to signals which are observable inside the factory. Such kinds of information, available even without communication with other firms, include own production of the firm in question, its sales, and its input and output stocks.

In the order-signal system decision-making is also decentralized. The situation is slightly different with information, because the flows of information are based on the order, l_{ij}, which is nothing else but *communication*

[18]This error of mine is also shared by those representatives of the Clower–Barro–Grossman school who like to give the name "disequilibrium theory" to their results. And yet a truly justified name would be the "theory of non-Walrasian equilibria".

between two decision-makers. In the present case it is *horizontal* communication: firms placing orders and those accepting and fulfilling them are at the same level as far as their legal status is concerned. (Although, as has been seen, the seller is the more powerful in a shortage economy.) It is worth remarking that at this point a special kind of horizontal communication is involved: the flow of information is linked to the physical flow of products.[19] Information flows are decentralized; buyer–seller pairs enter into informational contacts separately.

The theoretical propositions based on the model of section 7.6 indicate that – under a set of specific simplifying assumptions – a *perfectly decentralized abstract system can operate without price signals*. This leads on to far-reaching general issues. Theories of system comparison usually contrast two formations: decentralized control with price signals and centralized control without price signals but with "quantity" signals. We have shown that there exists a further variant, namely decentralized control with nonprice "quantity" signals. And this not only appears in the sphere of abstract models, but exists in reality and, consequently, is empirically observable. Decentralized "quantity" control plays a particularly important role in the resource-constrained shortage economy, but is to be found in other systems as well. This type of control is called the *vegetative control** of economic processes, by analogy with the special role played by the vegetative nervous system in the functioning of higher organisms. *Vegetative mechanisms control the most simple, elementary, and trivial adjustment of real economic processes.*

Of course, their role is not exclusive; they do not fulfill alone all the functions of control. But alongside other control mechanisms (supporting, or, on the contrary, in conflict with them) vegetative mechanisms also play a very important role.

While stressing that vegetative control always asserts itself, even in the traditional, strongly centralized form of economic management of the socialist system, we do not wish to attribute to it a greater role than it plays in reality. It is a well-known fact that in traditional socialist economic management, multilevel vertical control is prevalent, involving vertical communication. Downwards are sent mainly plan directives, and upwards

[19]Many other kinds of communication take place in reality. Not only between seller and buyer as in the case of orders, but also between two sellers (cartel agreements on prices, or in the division of markets).

*The term *autonomous control* is sometimes used in the literature to denote this concept. In the present book, however, we shall avoid confusion by adhering to the expression introduced in the text. (*Editor's note.*)

suggestions, applications, and reports about the fulfillment of directives. We shall supplement this familiar picture with a few remarks.

Some of the signals aiming at the center are – as we mentioned in Chapters 3 and 5 – quite closely related to those involved in vegetative control. "The queue has grown unusually, unbearably..." – that is a signal which sooner or later causes a reaction on the part of the central planners. And, conversely, they will also respond with an appropriate reallocation, to reports indicating that slack, stock, or unutilized capacity have accumulated somewhere.

The bulk of both downward- and upward-flowing information consists of "quantity" signals. This is worth emphasizing all the more as the practice of traditional socialist economic management system differs from much of the related theoretical literature in just that respect. Models of multilevel control usually describe a vertical exchange of information in which "quantity" information flows in one direction, and price information in the other.[20] This special kind of "duality" does exist in reality, but only within a rather narrow sphere.

There is no clear cut separation between vertical and horizontal flows of information. Central plan directives are often given in a relatively aggregated form, leaving disaggregation to direct negotiation between buyer and seller. In other cases central directives just sanction formally the previous agreements between the parties.

In the *postreform* Hungarian economic management system the vertical flows of information are much diminished (although they have by no means ceased). At the same time, the effect of price signals has remained rather weak in the enterprise sector – as is noted in Part II of the book. Precisely for this reason the role of vegetative control has grown.

7.9. Aggregation and "condensation" of market shortage and slack indicators

As with the subjects treated so far in the book, again we make a few remarks about problems of observation and measurement. I would like, however, to avoid repeating myself. Sections 2.10 and 2.11 discussed in detail the indicators describing shortage and slack in *production*. Section

[20] We have in mind the studies concerned with the economic interpretation of so-called decomposition algorithms, interpreting these not only as computational procedures but as *descriptive* models of multilevel control. See, among other works, Kornai–Lipták (1962, 1965), Kornai (1973, 1975a), Malinvaud (1967), and Heal (1973).

18.7 will discuss indicators suitable for measuring shortage *in trade between firms and households*. The methodological points made in these two places are also applicable, *mutatis mutandis*, for measuring shortage and slack in the sphere discussed in the present chapter, namely in interfirm trade.

Let us assume that we have available a number of indicators z and q which describe shortage phenomena on each interfirm partial market (insatiable initial demand, forced substitution, length of queues, search, etc.), as well as slack (rate of turnover of stocks, distribution of frozen stocks, unused service capacities, etc.). Yet the huge volume of partial data cannot be assimilated directly, either for the preparation of practical economic policy decisions, or for scientific analysis. The data must be "condensed" in some way, but before setting forth recommendations, I wish to criticize the forms of "condensing" that I consider mistaken.

The state of a shortage economy cannot be described summarily by the category of "aggregate excess demand". Those applying mechanically the neoclassical and Keynesian conceptual apparatus for the description of the socialist economy mostly apply this concept.[21] Yet this category is not applicable for the analysis of a system in which (1) shortage phenomena are *chronic* and the behavior of the participants of the system has permanently adjusted itself to shortage, and (2) the role of money is passive in a considerable proportion of the economy. Of course, I am not against all forms of aggregation, as is obvious from what has gone before; and I will shortly return to it. I only reject mistaken aggregation, and it is advisable to avoid the following errors.

Purchases in line with initial demand should not be added to purchases based on forced substitution. The buyer's total expenditure obtained in such a way should in no way be identified with "aggregate demand". It is totally erroneous to believe, if the difference between total income and total expenditure, that is total saving, does not show any rapid growth, that it proves aggregate demand has been satisfied. Not at all. It only shows that income owners have spent their money. A considerable part of their *initial* demand may not have been satisfied at all, and money intended for certain purchases might have been absorbed by forced substitution or other unintended spending. Distinguishing between purchases according to initial demand and those deviating from it may be very difficult, and cannot be done with any exactitude. Yet the roughest of estimations, or some indirect observations assisting such a distinction, would be worth more than any aggregation of items of totally different content. This is one

[21] See, for example, Barro–Grossman (1971, 1974), Portes–Winter (1977a, 1977b, 1978) and Howard (1976).

of the most fundamental mistakes of the Clower–Barro–Grossman school in its analysis of the socialist economy.

Shortage and slack must not be balanced out. Shortage in one sphere is not compensated by slack in some other sphere. "Excess demand" in one partial market and "excess supply" in another do not "net out" each other. The same is also true in a dynamic sense. Given unused railway capacity in the spring and congestion in the autumn one cannot say that utilization was just right "on the average". Shortage and slack phenomena must be recorded separately.

"Aggregate excess demand" is a *cardinal* indicator expressing the "volume" of shortage as a scalar magnitude. Even if that is impossible to determine, could we not at least find an indicator function of *ordinal* character as an overall expression of shortage? Mathematically it would not be especially difficult to specify an *index function* Z:

$$Z=f(z_1,z_2,\ldots), \tag{7.9}$$

in which Z increases monotonically in every argument. That is to say, if the intensity of shortage increases in any dimension, as measured by any indicator, while the value of the other indicators remains unchanged, the composite indicator Z would also show an increasing general intensity of shortage. The problem requires further research.

Let us acknowledge for the time being that shortage and slack are *vector* categories. "Condensation", that is a compact survey of the large number of indicators, can be carried out within the limits customary and permissible in vector analysis.

(1) The simplest procedure is to select a subset of the many thousands of potential indicators to represent the whole set.

(2) The aggregation of one or other indicator or the computation of averages for certain groups of products may be carried out. (For example, the forced substitution ratio for all building materials.) In order that a comparatively high degree of aggregation should not hide possibly very worrying shortages in limited areas, it should be supplemented by representative disaggregate indicators.

(3) The use of *dominance criterion* can be attempted. For example, a comparison may be made between the long periods (t_1,t_2) and (t_2,t_3) to find out whether the following conditions are fulfilled:

$$z_{ij}^*(t_1,t_2)\geq z_{ij}^*(t_2,t_3), \quad \text{for every product } i \text{ and indicator } j, \tag{7.10}$$

that is the intertemporal average of the shortage intensity did not increase for a single product, measured by any of the indicators, and

$$z_{ij}^*(t_1,t_2) > z_{ij}^*(t_2,t_3), \quad \text{for at least one } i \text{ and } j, \tag{7.11}$$

that is the intertemporal average of the shortage intensity definitely diminished for at least one product, measured by one indicator. In this case the vector z^* of the later period *strictly dominates* the vector representing the earlier period. We can then make with assurance a summary statement, concerning all markets and all production in the economy, that the intensity of shortage has diminished. And, of course, such a statement can be made particularly confidently if condition (7.11) is strongly fulfilled: intensity of shortage has diminished not only for one product but for several, measured not only by one but by several indicators. And, what is more, not only by 1 or 2 percent, which could be within the range of measurement errors, but to a considerable extent, eliminating any doubt that the shortage situation has changed.

For practical purposes it is not important to insist rigidly on a strict assertion of the first condition (7.10) of the dominance criterion. If the intensity of shortage has increased for comparatively few products, perhaps measured by only one or the other indicator, and to a small extent, while it has significantly decreased for some other indicators, we can still say that the dominance criterion *approximately* asserts itself. This "loosening", however, must be handled with care. It should rather be a reasonable acknowledgement of the uncertainties and inexactitudes of measurement. A 1 or 2 percent growth in intensity could be explained by such inexactitudes, so one would hardly refuse to apply the dominance criterion for that reason. But we must not go farther, otherwise we would return to the theoretical difficulty which was mentioned earlier in relation to the composite indicator function (7.9).

(4) It can be expected that the dominance criterion should be fulfilled comparatively rarely (strictly or even just approximately) for the entire vector. On the other hand, it can turn out that these criteria may be applied to a definite part – clearly defined by their economic content – of the vectors z and q. As has been mentioned, the following can certainly be said about the Hungarian economy. Comparing the periods 1949–53 of taut growth and 1968–72 following the reform, it is found that shortage intensity clearly diminished on the market and in the production of producer's goods, while it conspicuously grew on the labor market. It could be demonstrated for each sphere separately, how the dominance criterion

is asserted (at least approximately) with the inequality in the appropriate direction.[22]

This brings us as far as we can go in analyzing aggregation and "condensation" based on the vectors z and $\boldsymbol{q}$. Usually no fully comprehensive statements can be made covering all markets and the whole of production. We have to go at least one degree deeper and break down the whole into large parts. These parts can then be characterized by explaining whether their normal states remained unchanged or experienced long-term shifts. They can also be described by the departures of the instantaneous situation from the normal state, and by the volatility or relative stability of their actual states.

[22]In the Hungarian Press or economic policy speeches and public decisions the following phrase is often met with: "the *equilibrium situation* in some area of the national economy has improved (or worsened)". The mode of expression is somewhat strange, but whoever is acquainted with this language understands the meaning. It is, in our terminology, that the shortage intensity has decreased (or increased) in the area in question, and in the domination sense just explained, strictly or at least approximately. The usage of the technical jargon indicates that it is not only theory but also practice that requires *scales* up on which to measure shortage intensity, and "condensation" to aid the summary description of the shortage situation.

CHAPTER 8

Frictions of adjustment

8.1. Introduction

In Chapter 7 we changed the "mood" to some extent: the description of phenomena became more abstract as compared with the level of abstraction in Chapters 2–6. We now make further steps in the same direction. Since we want to analyze adjustment frictions on a theoretical level, but without the discussion becoming overcomplicated, we must make some extremely strong simplifying assumptions.[1] We shall carry out a few mental experiments, as it were under "laboratory conditions".

We will not define the concept of friction at this stage. In sections 8.2–8.6 we present a few examples through which the meaning of this important concept will gradually evolve. A general and summary discussion will then follow at the end of the chapter.

Most of what we have to say in this chapter is *not system-specific* but is equally valid for any system. It is only towards the end of the chapter that we discuss questions concerned with the friction present in the resource-constrained system.

As regards *temporal* aspects of the analysis, we shall discuss *comparative statics* in sections 8.2–8.4, section 8.5 deals with the buyer's *instantaneous* adjustment, and section 8.6 with *short-term* adjustment processes.

In sections 8.2–8.6 we consider *submicro-level* elementary events, and we develop some proportions about the *overall* system in the concluding sections of the chapter.

[1]There would be no obstacle to further developing the models of this chapter in order to remove a number of restrictive assumptions. However, the problems would become much more complicated mathematically. At this point it is not worth doing this. For the purpose of "warming-up" and as a first acquaintance with the problems, the present extremely simplified models seem to be adequate.

8.2. Relationship between stock and forced substitution

It has already been pointed out several times in the preceding chapters that shortage and slack may exist together, and what is more, that there are also *causal relations* between the two groups of phenomena. Let us examine a single example of these relations with the aid of a model.

We examine a market on which a single group of products is traded. There are m individual products belonging to the group, with certain quality differences between them, but able to substitute for one another. In our standard example, this could be the product group consisting of steel of qualities 9, 10, and 11. There is one seller (a monopolist, or an aggregate of sellers) and one buyer (similarly a monopolist buyer or an aggregate of buyers). Only one buying act is observed, and the observation is static. The following notation will be used:

$$d_1,\dots,d_m = \text{initial demand}; \qquad \sum_{i=1}^{m} d_i = D = \text{total initial demand};^2$$

$$s_1,\dots,s_m = \text{supply}; \qquad \sum_{i=1}^{m} s_i = S = \text{total supply},$$

$$y_1,\dots,y_m = \text{actual trade}; \qquad \sum_{i=1}^{m} y_i = Y = \text{total trade}.$$

If the buyer does not find the product corresponding to his initial demand, he is willing to accept any forced substitution within the same group of products.

The seller knows D, the total initial demand, exactly, and he also knows that the buyer is willing to accept any kind of forced substitution.[3] On the other hand, he has no information at all about the composition of initial demand. He applies the following *stock formation rule*:

$$s_i = (1+\lambda)D/m, \qquad \lambda \geqq 0, \qquad i = 1,\dots,m. \tag{8.1}$$

The logic of the rule is as follows. Since the seller does not know anything about the composition of initial demand, all m products will be represented in equal shares in his total supply. He wants to guarantee satisfaction of total demand, therefore the lower limit on the *stock factor* $(1+\lambda)$ is 1. If the *stock parameter* λ is positive, a *buffer stock* is also formed, in addition to the minimum stock, $S = D$. This would already be capable of

[2] Aggregation may be done in physical units of measurement or at fixed prices.
[3] We assume that $S \geqq D$.

satisfying, partly or fully, some components of initial demand, even if the share of these is not the same for the different products. The stock parameter, λ, is the decision variable of the seller. In our introductory example this is a *slack indicator*: the larger λ, the larger the slack on the market.

And now let us go on to describe the state of the buyer: his success or failure on the market. *Purchase based on initial demand* is denoted by $y^{(d)}$:

$$y_i^{(d)} = \begin{cases} d_i, & \text{if } d_i \leqq s_i \\ s_i, & \text{if } d_i > s_i. \end{cases} \tag{8.2}$$

The buyer tries to satisfy his initial demand. He goes up to the limit of initial demand or of supply. That part of his initial demand which still remains unsatisfied, he covers by *forced substitution*:

$$Y^{(FS)} = D - \sum_{i=1}^{m} y_i^{(d)}. \tag{8.3}$$

Let us denote the forced substitution ratio by η:

$$\eta = Y^{(FS)}/Y, \qquad Y = D. \tag{8.4}$$

The value of indicator η is 0 if no forced substitution was effected; it is 1 if purchases were all completed by forced substitution. (If the rule (8.1) is fulfilled, η cannot reach the value of 1, but may be $(1-1/m)$ at the highest.)

Since we do not know the composition of the buyers' initial demand, we cannot determine the value of η exactly. We can, however, provide an upper limit. We can determine the forced substitution ratio which the buyer will not exceed even in the case of the most unfortunate deviations between the composition of demand and supply. Let us call this the *maximum value* of the forced substitution ratio, denoted by $\hat{\eta}$:

$$\eta(\boldsymbol{d}, \boldsymbol{s}) \leqq \hat{\eta}, \quad \text{for any } \boldsymbol{d}, \boldsymbol{s}, \quad D \leqq S. \tag{8.5}$$

The maximum value of the forced substitution ratio[4] appears in our model as a *shortage* indicator. Although, in accordance with what was said in earlier chapters, the intensity of shortage should in fact be described by a vector, an ensemble of several indicators; however we shall use here, for simplicity's sake, only one. What is more, even this single indicator, $\hat{\eta}$, only indicates the magnitude of shortage in a roundabout way. What it indicates is not only what the actual *ex post* ratio of forced substitution was.

[4] $\hat{\eta}$ is the smallest of the upper bounds of $\eta(\boldsymbol{d}, \boldsymbol{s})$.

(This could be given by observation of $\eta(t)$ and by computation of its intertemporal average.) Indicator $\hat{\eta}$ gives information *ex ante* about the "most pessimistic" ratio, which is certain not to be surpassed by the actual forced substitution ratio. (It is another matter that, in a more fortunate case, the actual forced substitution ratio may be lower than this pessimistic ratio.)

The interrelation between the slack indicator λ and the shortage indicator $\hat{\eta}$ is shown in figure 8.1. The following proposition holds:

$$\text{If } \lambda = 0, \text{ then } \hat{\eta} = 1 - 1/m. \tag{8.6}$$

Let us assume that the total initial demand is concentrated on a single product. In this case a fraction $1/m$ of the demand can be satisfied in line with the initial intentions; the rest must be covered by forced substitution. The larger the number of products, m, the closer this extreme case is to 1, that is to the case where total purchases are based on forced substitution. And now let us consider the other extreme case:

$$\text{If } \lambda = m - 1, \text{ then } \hat{\eta} = 0. \tag{8.7}$$

Obviously, if the seller is willing to keep as much stock of each product as the total initial demand, the initial demand vector $d_1, d_2, \ldots, d_m$ can always be satisfied, even if all the initial demand is concentrated on a single

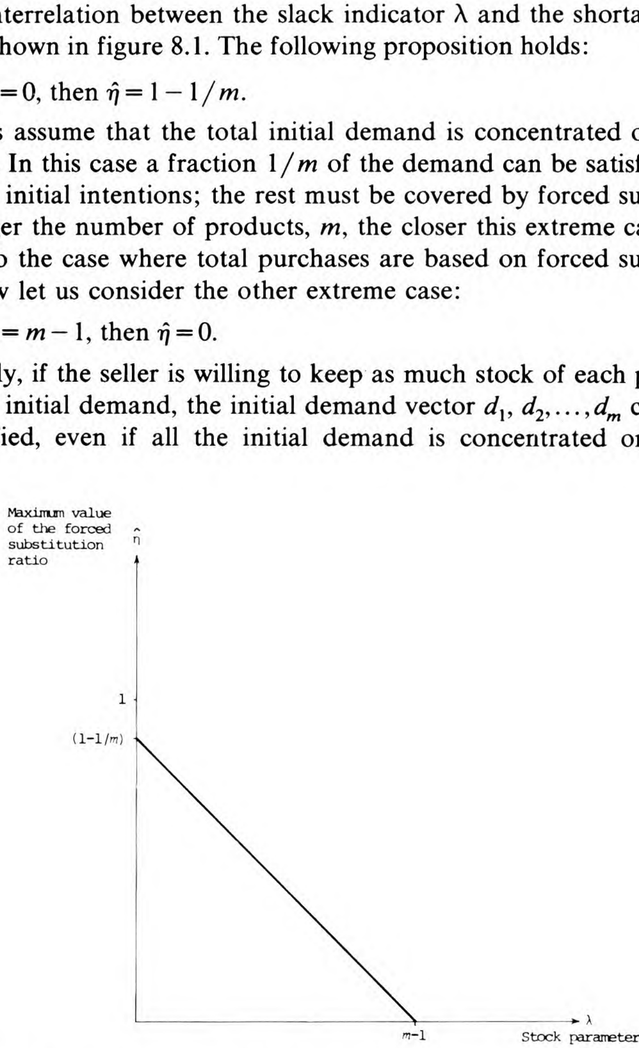


Figure 8.1. Relationship between the forced substitution ratio and the volume of the seller's stock.

product. It would be useless to keep any more stock than that, so we need not deal with the case $\lambda > m-1$.

If $0 < \lambda < m-1$, the maximum value of the forced substitution ratio will be on the straight line between the two extreme points.

Thus, we can conclude that *shortage intensity is a decreasing function of productive slack.* In our example the nature of this dependence is simple, the shortage indicator $\hat{\eta}$ being a decreasing linear function of the slack indicator λ. Of course, this mathematical form only holds under certain conditions; this is, however, of no particular importance. What really matters is the economic content of what we have to say: the negative relationship between slack and shortage. This is verified by countless empirical observations. It is self-evident for every buyer that if the seller keeps a stock of greater assortment and larger volume, he will have a better chance of receiving what he is looking for. If, in the case of several sellers, each holds larger stocks, or, in the case of services, the reserve capacity is larger, there will be a greater probability that any particular buyer can avoid forced substitution, search, queuing, or postponement of purchase.

In our example every stock policy $\lambda \leqq m-1$ leads exclusively to a *productive* slack, since every item of stock has some chance of being bought. This does not mean, however – as was explained in introducing the concept of "productive" versus "unproductive slack" – unconditional approval of the stock policy. Is the slack thus built "good" or "bad"? The answer depends on a value judgement about the importance attributed to the buyer's satisfaction. Let us assume that the seller accumulates the largest reasonable stock, that is $\lambda = m-1$. A large slack then results, since the buyer only buys a fraction of the available supply. How are we to evaluate the size of this slack? If we deem it important that the buyer should never be disappointed, the entire slack can be regarded as a *reserve*. Within the scope of the present model (total lack of differentiated information) any element of the stock has some chance of being bought. If, however, we do not mind if the buyer can hardly obtain anything in line with his initial demand, most of the slack in question is *superfluous*.

An objection may be raised against this train of thought that the issue does not appear so sharp in reality. We have assumed here that the seller has no idea at all of the expected composition of demand, though usually he can make reasonable forecasts. That is true, but, for the sake of sharper analysis, the extreme form of the model can be very useful. Whether or not there are reliable predictions of the buyer's expected demand, the importance of the buyer's satisfaction or disappointment continues to be a component of the decision problem. How important is it for the seller – or for society?

Of course, we do not thereby intend to avoid the study of demand predictions, and we shall now go on to examine this question.

8.3. Prediction error of the seller

In section 8.2 it was assumed that the seller knew nothing about the composition of the buyer's initial demand. In reality, of course, he would accumulate a lot of experience in observing the past behavior of the buyer and he would make use of this information in forming his supply.[5] Let us denote by $(d_1^{\text{pred}},\dots,d_m^{\text{pred}})$ the seller's *prediction* of the buyer's initial demand. As in the model of section 8.2 the seller knows here, too, the buyer's total initial demand and adjusts his prediction to it:

$$\sum_{i=1}^{m} d_i^{\text{pred}} = D. \tag{8.8}$$

The seller reckons that there will be some initial demand for every product:

$$d_i \geqq \delta > 0, \qquad \text{for every } i. \tag{8.9}$$

Of course, the prediction may be erroneous; it may be different from the buyer's true initial demand:

$$\varepsilon_i = \begin{cases} d_i - d_i^{\text{pred}}, & \text{if } d_i^{\text{pred}} < d_i: \quad \text{underestimate} \\ 0, & \text{if } d_i^{\text{pred}} \geqq d_i: \quad \text{correct prediction, or} \\ & \qquad\qquad\qquad\quad \text{overestimate.} \end{cases} \tag{8.10}$$

The variable ε_i measures the *under*estimation.

The total of the seller's errors of underestimation is denoted by ε. This is an *indicator of prediction error*:[6]

$$\varepsilon = \sum_{i=1}^{m} \varepsilon_i. \tag{8.11}$$

[5]The analysis carried out in sections 8.3 and 8.4 is based on research work carried out with András Simonovits. Here, in the following two sections, we expound only the most important ideas, with the aid of the simplest possible (and therefore rather special) models. We did not even try to set out in detail every implied assumption. A fuller account is in Mathematical Appendix B where we present and analyze two models of somewhat more general character; the models of sections 8.3–8.4 are special cases of these more general models. Also, Mathematical Appendix B provides the mathematical proofs of all propositions.

The same problems are examined in the study by Kornai–Simonovits (1977b) using other types of model (e.g. dynamic models, learning processes).

[6]For assumptions underlying the formation of indicator ε see section B.1.

If there is no difference between the true initial demand and the seller's prediction, in any component, then $\varepsilon = 0$. The maximum value of the indicator of prediction error of the seller, $\varepsilon^{\max} = D - \delta$. This arises if the buyer's entire initial demand is directed to one favored product, yet the seller predicted the demand for this product to be the minimum value δ.

In this modified model the seller applies a stock formation rule somewhat different from formula (8.1):

$$\underbrace{s_i}_{\boxed{\text{supply}}} = \begin{cases} \underbrace{(1+\lambda)}_{\boxed{\text{stock factor}}} \; \underbrace{d_i^{\text{pred}}}_{\boxed{\text{seller's demand prediction}}}, & \lambda > 0, \quad \text{if } \underbrace{(1+\lambda) d_i^{\text{pred}} < D}_{\boxed{\text{supply smaller than total demand}}}, \\ \underbrace{D}_{\boxed{\text{total demand}}}, & \text{otherwise}. \end{cases} \tag{8.12}$$

The logic of the rule is as follows. The seller's starting point is not quite as in section 8.1: there, not knowing anything about the composition of demand, he let each product take an equal share in the stock. His starting point now in his own prediction about the composition of demand. This may, however, be "multiplied" by the stock factor $(1+\lambda)$. If he does not entirely trust his own prediction, he may wish to be on the safe side, and accumulate additional supplies, so $\lambda > 0$.

In setting up formula (8.12) we made an effort to provide the simplest possible rule of thumb. The rule oversecures in fact. It can be demonstrated that in the case of applying a more complicated rule a smaller stock may be sufficient to satisfy demand with the same security.

Under such circumstances the maximum value of forced substitution is not a function of one variable as in section 8.2, but of two variables:

$$\underbrace{\hat{\eta}}_{\boxed{\text{maximum value of the forced substitution ratio}}} = \phi(\underbrace{\lambda}_{\boxed{\text{stock parameter}}}, \underbrace{\varepsilon}_{\boxed{\text{seller's total prediction error}}}). \tag{8.13}$$

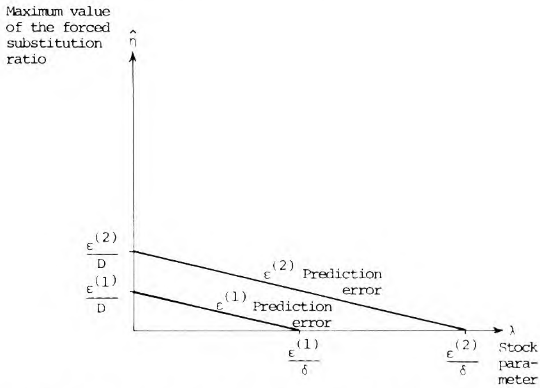


Figure 8.2. Relationship between the forced substitution ratio, stock volume, and prediction error.

The function (8.13) is illustrated in fig. 8.2.[7] The stock parameter is measured on the horizontal axis and the maximum value of the forced substitution ratio on the vertical axis. The iso-lines represent $\hat{\eta}$ as a function of λ, for fixed levels $\varepsilon^{(1)}$ and $\varepsilon^{(2)}$ of the prediction error.

Let us denote the upper limit of the stock parameter by $\hat{\lambda}: \hat{\lambda} = \max(D/d_i^{\text{prod}} - 1)$. It is certain that at $\lambda = \hat{\lambda}$ demand with any composition would be satisfied. If the stock parameter exceeded the upper limit $\hat{\lambda}$, such slacks would accumulate as had no chance of being bought, that is *unproductive* slacks. If the parameter λ reached the value $\hat{\lambda}$, supply of size $s_i = D$ of each product would be available in accordance with formula (8.12). This guarantees the total elimination of forced substitution even when there are prediction errors. The result can also be worded in the following way: *even the worst prediction can be compensated for by increasing slack.*

The upper limit of the maximum value of the forced substitution rate is $(D - \delta)/D$. This follows from the stipulation made about the prediction of demand that a stock of at least δ of each product must be held.

The maximum value of the forced substitution ratio will be zero if $\varepsilon = 0$. In this case[8] there is no need for a buffer stock (larger than $S = D$), so $\lambda = 0$. In other words, *in the case of perfect foresight Walrasian equilibrium may be achieved*: initial demand and supply agree perfectly. In our model $\hat{\eta}$

[7]More exactly, fig. 8.2 represents only a special case, namely, a market on which there are only two products, and the demand for one of them was underestimated by the seller. The more general problem is discussed–together with complementary assumptions not specified here–in section B.1 of Mathematical Appendix B. Function (B.12) is a specific form of relationship (8.13), valid with the assumptions described in section B.1.

[8]This case may only occur if the condition that true initial demand, $d_i \geqq \delta$ also holds.

represents the shortage vector z and λ represents the slack vector $\boldsymbol{q}$. Within the framework of our model the condition for Walrasian equilibrium can be formulated in the following way:

$$\underset{\boxed{\text{actual trade}}}{\boldsymbol{Y}} = \underset{\boxed{\text{initial demand}}}{\boldsymbol{d}} = \underset{\boxed{\text{supply}}}{\boldsymbol{s}}, \quad \text{i.e. } \underset{\boxed{\text{no shortage}}}{\boldsymbol{z}=\boldsymbol{0}}, \quad \underset{\boxed{\text{no slack}}}{\boldsymbol{q}=\boldsymbol{0}}, \text{ if } \underset{\boxed{\text{no prediction error}}}{\varepsilon=\boldsymbol{0}}. \tag{8.14}$$

This is the origin in fig. 8.2. Every market on which the seller's prediction is imperfect is situated away from the origin, in the region with positive values of the three variables.

The mathematical form of function (8.13) is, of course, special. The extremely simple form (linearity, etc.) illustrated in fig. 8.2 only holds within the framework of our model. Yet the economic content of the function illustrates a relationship of general validity. *Intensity of shortage is larger, the smaller the productive slack and the larger the error in the decision-maker's predictions.* Three groups of phenomena, namely shortage, slack, and prediction, have been presented here in the simplest possible form and each represented by a real number. Recognition of this "triple interrelation" is of basic importance in understanding the adjustment properties of economic systems.

Mutual trade-offs exist between the three variables of the relationship. *If prediction improves, less production slack is required for the shortage intensity to remain at the same level. Or, if we want to fix productive slack at a constant level, improvements in prediction will reduce the shortage intensity.*

As a consequence of the seller's demand prediction error a *friction* appears in the adjustment process. The variable ε is a *friction factor* in the system.

In what follows, interrelations between shortage, slack, and friction will be called *friction functions*. The first example of this is the relationship (8.13), a function of two variables: $\hat{\eta}=\phi(\lambda, \varepsilon)$.

8.4. The vacillating buyer

Let us now look at another type of adjustment friction, caused by *the volatility of the buyer's intentions.*[9] While repeating purchases, his initial demand may be subject to fluctuations as a result of various factors. A few examples in the sphere of the firm's material demands may be given.

[9] A similar problem is treated in the study by Gordon–Hynes [1970]. The authors point out that in the case of stochastic demand, "excess supply" cannot be zero; instead, the seller

(a) The production plan of the firm and its composition by product may vary. Input requirements may fluctuate accordingly.

(b) One or other kind of material is in most cases only utilizable together with other complementary inputs. If the supply of complementary inputs fluctuates, demand for the material in question may fluctuate in parallel.

(c) The planning activity of the firm is superficial, its management disorganized, with the result that claims are submitted capriciously.

(d) As a result of innovations or new technologies new products are introduced which may suddenly change input requirements.

We do not form any value judgement about the volatility of buyer's intentions. In cases (a) and (b) it is explained by circumstances outside the firm; in case (c) it is the firm's own fault, and in case (d) it is a welcome change concomitant with innovation. We are now simply stating that the buyer's intentions may vacillate. For an analysis, we return to the model described in section 8.2, except that the buyer's demand is now regarded as a stochastic phenomenon:

$$\underbrace{d_i}_{\text{initial demand}} = \underbrace{\bar{d}_i}_{\text{deterministic part of initial demand}} + \underbrace{\gamma}_{\text{vacillation parameter}} \underbrace{h_i}_{\text{random part of initial demand}}. \qquad (8.15)$$

The term $\bar{d}_i$ is the mean value of the fluctuating initial demand; it is deterministic. To this is added, or from it is subtracted, the stochastic part γh_i. The expected value of the random variable h_i is zero; thus the actual initial demand may deviate from $\bar{d}_i$ in both directions. Its probability distribution may be different for each product.[10] The random variable h_i is multiplied by the *vacillation parameter* γ: $\gamma \geqq 0$. If $\gamma = 0$, the buyer is not vacillating at all, and we have the deterministic model of section 8.2. If on

becomes interested in stockholding. In our own terminology, there is a relationship between "search" carried on by the buyer and the size of slack.

Lack of information and uncertainty also lead to similar questions in the labor market. (See Phelps, 1970b, and Holt, 1970.)

The approach used in the present chapter is related in a number of aspects to that of the volume edited by Phelps (1970a). The three works mentioned above were also published in this volume.

[10] The distribution is truncated, since the condition $\sum_{i=1}^{m} d_i = D$ holds and $d_i \geqq 0$ for every i. For the sake of simplicity we assume that the distribution is continuous.

the other hand, $\gamma > 0$, the buyer is vacillating. In our model we adopt the strong simplifying assumption that the parameter γ takes the same value for every product.

The seller does not know the distribution of the variable h_i. He goes on applying the stock policy described in the rule (8.1). Yet we, theoretical analysts of the problem, know the probability distribution and we are therefore able to make a better prediction of shortage intensity. As an indicator we again use an upper limit of the forced substitution ratio, but now with reliability π $(0 \leqq \pi \leqq 1)$. The indicator will be referred to as the *reliability value* of the forced substitution ratio, and denoted by $\eta^{(\pi)}$. Accordingly, if, for example, $\pi = 0.97$, there is a 3 percent probability that the *actual* forced substitution ratio will be larger than the reliability value $\eta^{(0.97)}$, and 97 percent probability that it will not be larger:[11]

$$\begin{aligned} &\eta \leqq \eta^{(\pi)}, \quad \text{with probability } \pi, \\ &\eta > \eta^{(\pi)}, \quad \text{with probability } (1-\pi). \end{aligned} \tag{8.16}$$

After outlining the model let us summarize briefly the main conclusions to be drawn from it.

We shall arrive at a "triple" interrelation similar to before, when discussing the seller's prediction error. This is shown in fig. 8.3.[12] As can be seen, it is very similar to fig. 8.2. Here, however, shifts of the iso-lines represent not changes in the seller's prediction error but in the buyer's vacillation. Mutual trade-offs exist between the three different groups of phenomena:

Shortage intensity will be the larger, the smaller the productive slack and the larger the buyer's vacillation. A larger extent of buyer's vacillation can be compensated by more productive slack. But if that is not done, the buyer will have to pay for larger vacillation with a more intensive forced substitution.

The similarity of the figures should not deceive us. The shifts of iso-lines in figs. 8.2 and 8.3 represent two different groups of phenomena. The latter

[11]Initial demand being stochastic, the forced substitution ratio is also stochastic. Let χ denote the probability distribution of the random variable η. The indicator $\eta^{(\pi)}$ is the *quantile* of the distribution χ at the reliability level π.

If, after all this, we return to the deterministic version of the model, we can also interpret definition (8.5) as a limiting case of (8.16) for the case $\pi = 1$.

[12]The figure presents a linear relationship. This is valid if the probability distribution of the random variables is uniform. This assumption is used only to simplify the figure. The economic content of what we have to say also holds for other distributions and, accordingly, for the case of nonlinear interrelations as well.

Complementary assumptions serving as background to the figure are described in more detail in section B.2, Mathematical Appendix B.

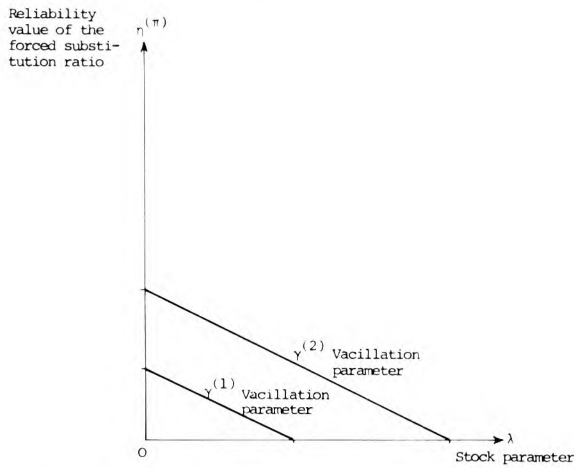


Figure 8.3. Relationship between the forced substitution ratio, stock volume, and buyer's vacillation.

expresses fluctuation of the buyer's actual behavior, while the former expresses errors in the "image" formed in the seller's mind about the buyer's behavior. Yet, in spite of the differences, they are also closely interrelated. *Both cause friction in the adjustment processes.* Like the earlier variable ε, the variable γ describes a *friction factor*. In accordance with this, we can formulate a new *friction function* in two variables:[13]

$$\underbrace{\eta^{(\pi)}}_{\text{reliability level of the forced substitution ratio}} = \phi'(\underbrace{\lambda}_{\text{stock parameter}}, \underbrace{\gamma}_{\text{vacillation parameter of the buyer}}). \tag{8.17}$$

The present friction function ϕ' and the function ϕ mentioned in the previous section are closely related.[14] In both of them, friction variables ε and γ affect the other two variables, namely the indicators of shortage and productive slack in the same direction. Whether it is the seller's prediction

[13]Function (B.17) in Mathematical Appendix B is a specific form of relationship (8.17), valid with the assumptions described in section B.2.

[14]See the analysis of "equivalence relations" between friction functions in Kornai–Simonovits (1977b).

error or the buyer's vacillation that increases, shortage intensity (with a given slack) or slack (with a given shortage intensity) has to increase with it.

8.5. Imperfect information of the buyer

We now turn our attention to a third kind of friction phenomenon which is the consequence of the *buyer's imperfect information*. So far we have talked about one seller and one buyer. We now examine a market on which there are several sellers and several buyers. We have to introduce even more simplifications to avoid complicating the discussion too much. The problem is schematically represented in fig. 8.4, while indicators of the state of the market are summed up in table 8.1.

There are three products: 1, 2, and 3. There are three buyers, each with an initial demand one unit of product. Buyer A wants to obtain product 1, buyer B wants product 2, and buyer C wants product 3. There are three sellers, each with a supply of one unit of product. Seller F has product 1 in stock, seller G has product 2, and seller H has product 3.

The market functions as follows. Goods arrive at the seller's store at a given time, let us say Monday evening, after which there is no further delivery for several days. All three buyers set out shopping the following day, Tuesday morning. That day each of them visits only one selling place. If the situation requires the buyer to look further for the product, he can make his next attempt only on the following day, his third on the third day, and so on.

As we can see, the starting point is highly favorable: initial demand and physical supply on the market are equal, and not only in total, but also in detail, in terms of individual products.

And now let the buyers set out on their way. Many cases are feasible of which we examine a few.

Case I. Things turned out very fortunately; every buyer visited the seller who held in stock the product he required. As is clear from table 8.1 there is no shortage, and already by the end of day one there is no slack.

Case II. Buyer A was also fortunate in this case; at his first attempt he found and bought the product 1 that he wanted.

But buyers B and C went to the wrong place: buyer B visited seller H, and buyer C visited seller G. Let us recall the shopping algorithm: in such cases the buyer may choose to take different steps. Of these only three will be examined now.

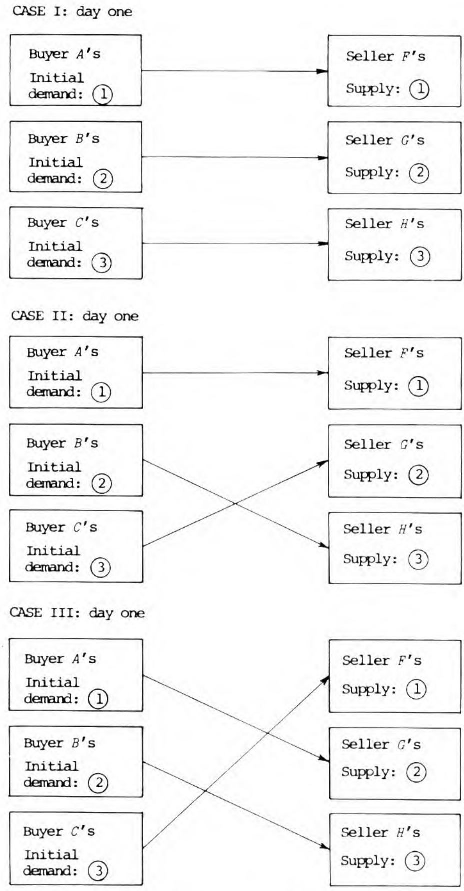


Figure 8.4. The buyer's initial route as a function of his information.

Table 8.1
Shortage and slack as a function of the buyer's information.

Indicator summarized for the whole of the market	Case I	Case IIa	Case IIb	Case IIc	Case III
Shortage					
Forced substitution ratio	0	2/3	0	2/3	1
Search	0	0	2	1	3
Slack					
Terminal stock at the end of day one	0	0	2	1	2
Terminal stock at the end of day two	0	0	0	1	1
Terminal stock at the end of day three.	0	0	0	0	0

Case IIa. Neither buyer is willing to search further, but applies forced substitution. The forced substitution ratio of 2/3 expresses the shortage. The decision of the buyers immediately absorbed the slack.

Case IIb. Both buyers are willing to search further. The following day they set out shopping again and have more luck, since they find the wanted product. In this instance the shortage indicator was search: summing over the whole market, there are two additional routes.[15] Slack was positive at the end of day one, since the supply of sellers G and H remained in stock. The following day, however, it was bought.

Case IIc. Buyer C is unwilling to search further and immediately carries out forced substitution. Buyer B does not know about this, but chooses to try again the following day, visiting seller G on that occasion. But there he finds empty shelves. He is forced to go back to seller H on the third day and accept forced substitution.

The summary result at the end of day three describing the state of the whole market, is as follows.

Forced substitution ratio is 2/3, the same as in case IIa. Yet the search indicator shows two additional routes; in this respect the shortage situation has worsened in comparison to case IIa. Meanwhile slack indicators showed unutilized stocks on two evenings, in contrast to case IIa, when slack was fully absorbed by forced substitution.

Although case II could have further subcases, we shall not continue with their description, but take a step forward.

Case III. All three buyers set out on the wrong route, and failed to find the desired product at the seller first visited. The story could have various kinds of continuation; let us examine one subcase.

Buyer A immediately applies forced substitution. Buyers B and C try again the following day and fail. Buyer B visited seller H and, not having found the required product, accepts forced substitution. Buyer C visited the same place and arrived some time later than B. He is distressed to hear that the desired product had been there but has been taken right from under his nose. The following day he visits seller F, where finally he does accept forced substitution.

The final balance of case III is very bad: the forced substitution ratio is 1. Altogether, three units of search have taken place: one additional route for buyer B and two additional routes for buyer C. Unsold stocks are considerable: two units at the end of day one, and one unit remained even after the second day.

[15]The search indicator for each buyer is just the number of selling places visited, minus 1. At least one seller must be visited in any case; this is not yet "search".

In table 8.1 search was summarized for the whole of the market. In case IIb we obtain the value of the search indicator, 2, by adding the one additional route of buyer B to the additional route of buyer C.

Although the story described above is very simple, it enables us to draw conclusions of some general validity. The fact that initial demand and physical supply are absolutely identical, $\boldsymbol{d}=\boldsymbol{s}$, implies the *possibility* of Walrasian equilibrium: $\boldsymbol{y}=\boldsymbol{d}=\boldsymbol{s}$; $\boldsymbol{z}=\boldsymbol{0}$, $\boldsymbol{q}=\boldsymbol{0}$. This was called here "*case I*". But what determines whether case I will occur, and not case II or case III, that is market states *deviating* from Walrasian equilibrium?

We could say that it is a matter of chance. In case I buyers were led by chance to the right selling place, and in other cases to the wrong one. Certainly chance *also* plays a role, but this cannot decide the question. As was said about the seller in the previous section, may now be said about the buyer: he has experience from earlier purchases, and now, too, he can obtain information before setting out on his shopping route. *The buyer's partial decision in the shopping process depend on his information.*

The pieces of information the buyer has about the seller's supply and about his fellow-buyers' attitude and actions may have many different compositions. We shall distinguish, for purposes of illustration, three different compositions, that is three different *information bundles*.

Information bundle I_0. All three buyers find out exactly which seller keeps what in stock. Thus, it is not by chance but knowingly that they can set out for the right place; case I can come about.

Information bundle I_1. This consists of two parts. *Part* (*a*): all three buyers know that product 1 is available from seller F, and also that a unit product 2 or 3 is supplied by seller G and H. But they do not know which of them has product 2 and which one product 3. *Part* (*b*): the other two buyers are also aware that buyer C's initial demand is directed at product 3. They also know that buyer C has a high forced substitution propensity and low search propensity. It is to be expected that if he does not find the wanted goods at the first attempt, he will immediately accept forced substitution.

Having all this information, the worst case III can be avoided even in the least fortunate circumstances. What is more, even the comparatively disadvantageous subcase IIc can be avoided. Depending on fortune, case IIb, IIa or I will come about.

Information bundle I_2. This is identical to part (a) of the bundle I_1 but does not include part (b). It is sufficient to ensure that buyers avoid the worst case III, but it does not exclude case IIc, which is the least satisfactory variant of case II.

The ordering of the three different information bundles suggests intuitively that I_0 is the "richest" in information, I_1 is "less rich", and I_2 is the "least rich". Formulated more exactly, we can construct a *partial ordering* on the set of alternative information bundles. Let there correspond to this ordering a *buyer's information imperfection* indicator Γ. The ordering and the corresponding indicator have the following properties:

$$\Gamma(\mathrm{I}') < \Gamma(\mathrm{I}''), \quad \text{if} \begin{bmatrix} z(\mathrm{I}') \\ q(\mathrm{I}') \end{bmatrix} \text{ dominates } \begin{bmatrix} z(\mathrm{I}'') \\ q(\mathrm{I}'') \end{bmatrix} \tag{8.18}$$

and

$$\Gamma(\mathrm{I}_0) = 0, \quad \text{if } z(\mathrm{I}_0) = \boldsymbol{0} \text{ and } \boldsymbol{q}(\mathrm{I}_0) = \boldsymbol{0}.$$

In accordance with this, the buyer's information imperfection indicator has been defined so that its value is zero when the buyer has *perfect information*. Its value for I′ is lower than that for I″, if I′ leads to less intensive shortage and/or smaller slack than I″ (based on the principles of vector comparison explained in the preceding chapter).

We cannot determine a perfect ordering over the set of alternative information bundles. If information bundle $\check{\mathrm{I}}$ does not dominate information bundle $\hat{\mathrm{I}}$, because, for example, it leads to less intensive shortage in respect of one phenomenon and to a more intensive one in respect of the other, we cannot state clearly which bundle is in fact richer in information.

If the buyer is more or less uninformed, the *value of information* becomes tangible. It is a well-known phenomenon in a shortage economy that the seller does a favor to the preferred buyer not by giving him the goods. (He is obliged to give them to anyone who asks for them.) The seller does a favor by *informing* the buyer that the required goods will arrive at such and such a time. Let us recall case IIc discussed just now. Buyer B visited seller G in vain, because buyer C had bought the goods before him. Yet if he had been told to hurry up, because the desired product was there, he could have been the one to get it. From this may be derived a measure of the value of information:

$$\boldsymbol{b}_i(\mathrm{I}', \mathrm{I}'') = z_i(\mathrm{I}'') - z_i(\mathrm{I}'). \tag{8.19}$$

$\boldsymbol{b}_i(\mathrm{I}', \mathrm{I}'')$	$z_i(\mathrm{I}'')$	$z_i(\mathrm{I}')$
value of additional information	shortage in case of less complete information	shortage in case of more complete information

The value of information is expressed by the diminution of the shortage experienced by buyer *i*: less search, less forced substitution, and so on. We do not give a scalar indicator of the value of information, but determine the difference between the two shortage vectors.[16]

[16]In addition to other differences, it differs in this from the definition of "information value" used, for example, by Marschak–Radner (1972). Otherwise our approach shows some similarities to that of Marschak and Radner.

We have now reached the threshold of an immense complex of problems: *the theory of economic information.* We have to stop at the threshold because, no matter how important the question of information is and how closely interrelated with the examination of shortage it is, its detailed discussion would lead us far beyond the scope of this book. I must content myself with formulae (8.18) and (8.19) and earlier (8.10) and (8.11), some simple examples indicating the *possibilities of measuring economic information.* There is no space to relate these special and simple formulae to the general theoretical framework for the measurement of economic information.

We now return to the original train of thought of this chapter. Indicator Γ further extended the list of *friction factors.* About this, the same thing can be said as about the other friction factors, namely that *the more imperfectly informed the buyer, that is the greater the friction in adjustment, the more intensive will be the shortage or the greater the slack.*

Our illustrative example involved considerable simplification, but the phenomenon itself is well known in everyday life. And it is particularly well known to buyers in the shortage economy. Complaints are often made because, the goods wanted are in fact available, although not at the place or time when they were sought. We do not say, of course, that this is the main reason for shortage. Earlier and later chapters stress that the main direct explanatory factor is the "running away" of initial demand, while, of course, physical supply is limited. The imperfect information of the participants in allocation processes may, nevertheless, amplify (and possibly cause in itself) shortage phenomena.

8.6. Delay and rigidity in the seller's adjustment

The following friction phenomenon is expressly associated with the *dynamics* of adjustment. We present the problem again via a very simple example, with numerous simplifying assumptions. A single seller is faced with a single buyer, just as in sections 8.2–8.4. Time t is an integer variable; its unit could be, for example, a week or a month. In any case, we have in mind *short-term* adjustment.

On this occasion there is no problem with information. The seller knows exactly that at a certain point in time denoted by t_0, the buyer's initial demand suddenly changed. Prior to this it had always been d; then it changed to $d' > d$ and from that time on the buyer always submits this new initial demand d'. Let us assume that there was no shortage before the critical date, t_0:[17]

$$\underset{\boxed{\text{supply}}}{s(t)} = \underset{\boxed{\text{initial demand}}}{d}, \qquad t<t_0. \tag{8.20}$$

The question is then how far supply is adjusted to meet the new initial demand? How long is the reaction, or adjustment period? The path of adjustment is represented in the following way:[18]

$$\underset{\boxed{\text{supply}}}{s(t)} = \underset{\boxed{\text{resistance parameter}}}{\Omega(t)} \quad \underset{\boxed{\text{old initial demand}}}{d} \quad +(1-\Omega(t))\underset{\boxed{\text{new initial demand}}}{d'}. \tag{8.21}$$

If the value of the *resistance parameter* Ω is 0, adjustment has taken place perfectly and supply has already adjusted to the new initial demand. If $\Omega=1$, no adjustment has taken place, and supply continues to correspond to the old initial demand. Many kinds of factor can explain resistance. A subjective resistance on the part of decision-makers may be observed. The most obvious explanation is the desire for a quiet life. It is always easier to repeat the old pattern of behavior while changing to something new requires effort and thought. If there is nothing to encourage additional effort and thought, why do it? Sometimes there are special interests motivating a cautious response. But beyond people's conscious or unconscious resistance, "things" resist in themselves. It is usually technologically impossible to change production or product supply from one moment to the next.

Let us denote by t_1 the time by which complete adjustment to the new initial demand takes place. Let us call the period $\tau=t_1-t_0$ the *delay time*, and denote it by τ.

We present a few characteristic illustrations of the dynamics of adjustment in fig. 8.5. In part A we have an extreme situation, where supply is

[17]For the sake of clarity let us assume that the change in initial demand took place at the last moment of week t_0. At the beginning of that week the buyer still appeared with his old initial demand, and the following week he presented his new demand.

[18]For simplicity, we consider the equality of supply and demand as full adjustment. The introduction of slack formation would cause no difficulty, but description would become more complicated.

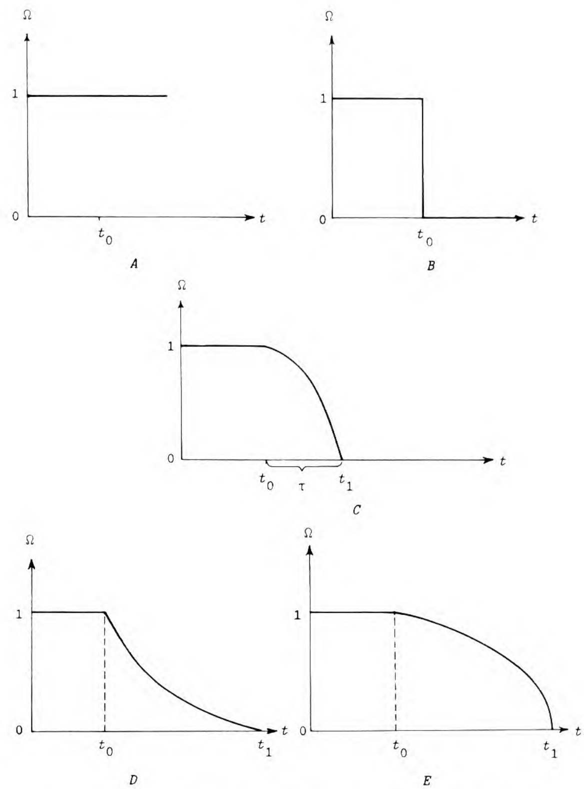


Figure 8.5. Characteristic types of delay and rigidity.

perfectly rigid. The seller stubbornly insists on supply corresponding to the old d. The resistance parameter, $\Omega(t)=1$ for every t, so that the delay time is infinite: $\tau=\infty$.

Part B presents the other extreme. There supply adjusts itself to the new initial demand without the slightest delay. There is no resistance, so $\Omega(t)=0$, and the delay time, $\tau=0$. This "pure case" is a mere abstraction, which cannot occur in reality.

There are innumerable intermediate possibilities between the two extremes, of which only three will be presented here. If we observed only the delay time, we would contrast case C with cases D and E, the delay times of the latter two being identical. In case C adjustment takes a short time, in cases D and E it takes longer.

The last two parts show clearly, however, that the delay time alone is not sufficient to characterize the dynamics of adjustment, since what happens between the beginning and end of the adjustment process is not a matter of indifference. A useful additional indicator is the intertemporal average of the resistance parameter:

$$\zeta = \frac{\sum_{t=t_0}^{t_1} \Omega(t)}{t_1 - t_0}. \tag{8.22}$$

Measured by this indicator it is clear that $\zeta^{(D)} < \zeta^{(E)}$. Obviously other indicators could also be used for characterizing the dynamics of adjustment; for present purposes, however, we can be content with the *rigidity indicator*, ζ. In other words, we shall say that the seller's behavior was more rigid in case D than in case E.

The above discussion has led us to two new *friction factors*, with two *friction indicators* to measure them, namely the delay time τ and the rigidity indicator ζ. These have formal properties similar to the ones mentioned earlier. In the case of perfect, frictionless adjustment $\tau = 0$, $\zeta = 0$. In reality the values of indicators is usually positive, a larger value indicating more friction.

These friction factors are connected to shortage and slack by a similar relationship, that is a *friction function*, similar to those involving the factors mentioned earlier. *The greater the lags and the more rigid the adjustment of supply to changes in the initial demand, the more intensive the shortage will be: the buyer is forced to search more, accept more forced substitution, and so on. And meanwhile, as a consequence of differences between physical supply and initial demand, a larger slack may develop.* And all this, for the friction forms just discussed, is not necessarily associated with any weakness in the *signalling system*. As was explained in Chapters 3, 6, and 7, there would be nonprice "quantity" signals which would direct the producer and the seller as to what changes have taken place on the demand side. This , however, is not enough. It is also important that supply should *react* flexibly and with little delay to the signals received.[19] If it were possible and worthwhile to establish an order of importance, we would regard this (i.e. friction of action) as the larger problem, and not the frictions in the signalling system, which were treated earlier in the present chapter.

[19]Lapan (1977a, 1977b) presents formalized models for the study of the phenomena called here *rigidity* and *delay*. Although the articles assume special institutional conditions, their results can be generalized.

8.7. First synthesis: The triple relationship of shortage, slack, and friction

Now the tools are all available for the first synthesis. In table 8.2 the friction indicators introduced so far are briefly surveyed. Columns II and III show the terms and notation introduced in sections 8.2–8.5. What is new is column I in which the individual frictional phenomena are formulated *more generally*. It can then be observed that those we discussed are special cases of more general phenomena. For example, in section 8.5 we analyzed the adjustment of supply to a change in initial demand. We could also speak of the symmetrical problem, namely the adjustment of initial demand to supply. In Chapters 4 and 5 this was treated in detail, though without stressing the frictions of this adjustment.

In column I of table 8.2 the expression, *actor*, occurs more than once. The term refers to any member of the system under consideration: individuals and groups (e.g. firms, households, nonprofit institutions). This is what replaces here the two different actors, seller and buyer, referred to in earlier parts of the chapter.

We do not claim that table 8.2 lists all the friction phenomena of economic adjustment processes or all the indicators which could be considered in their measurement; we are not even sure if those given are indeed the most important phenomena or the best measurement indicators. It is also possible that they should be classified differently. We hope neverthe-

Table 8.2
Friction indicators.

General contents of friction indicator	Special case presented in sections 8.2–8.5	
	Term	Symbol
I	II	III
The actor's imperfect information about other actors' state and attitude; error in predictions serving as a basis for actor's plan	Imperfect information of the buyer; the seller's demand prediction error	Γ, ε
Fluctuation of the actor's intention, in the case of repeated decisions	The buyer's vacillation	γ
Rigidity and delay in the actor's adjustment to changed conditions	The seller's rigidity; delay in adjustment to initial demand	τ, ζ

less to have made it clear in the course of the preceding discussion what we understand by friction, and the examples of indicators [Γ, ε, γ, τ, ζ] have confirmed that friction phenomena are measurable.

Friction indicators are brought together in the *friction vector* denoted $\boldsymbol{w}$, which includes in its components the friction indicators for all partial markets. For each partial market, the five indicators in table 8.2 (or other, more appropriate indicators in their place) are required. All components of the vector $\boldsymbol{w}$ have the formal properties that were already prescribed in the examples introduced in sections 8.3–8.6. That is, $w_h = 0$ means that the friction phenomenon measured by indicator h is not present at all; the system is free of friction from this point of view. The larger the value of w_h the larger the degree of friction in question.

If we observe friction phenomena over a long period, and calculate an intertemporal average, we can determine the *normal friction* $\boldsymbol{w}^*$ of the system.[20]

We can now discuss the "triple relationship" of shortage, slack, and friction.

Let us represent the shortage indicators of production and trade in the vector $\boldsymbol{z}$.

Let us represent the indicators of productive slack present in production and trade in the vector $\boldsymbol{q}^{\text{prod}}$. As a reminder, note that productive slack is that which *ex ante* has some chance of being used. In addition, there may also be unproductive slack, and $\boldsymbol{q} = \boldsymbol{q}^{\text{prod}} + \boldsymbol{q}^{\text{unprod}}$.

Finally, let the vector $\boldsymbol{w}$ represent the indicators of friction present in production and trade.

The following relationship obtains between the normal values of these variables:

$$\boxed{\begin{array}{l}\text{the smaller } \boldsymbol{q}^{\text{prod}*}\text{, i.e.}\\ \text{the normal value of}\\ \text{productive slack and}\\ \text{the larger } \boldsymbol{w}^*\text{, i.e. the}\\ \text{normal value of friction}\end{array}} \longrightarrow \boxed{\begin{array}{l}\text{the larger } \boldsymbol{z}^*\text{, i.e.}\\ \text{the normal value}\\ \text{of shortage in-}\\ \text{tensity}\end{array}}. \tag{8.23}$$

This is called the *general friction function* of the system. We have drawn it up in verbal form, because – at this level of generality – we have not yet arrived at a more exact clarification of the interrelation. Only a few

[20] All the methodological remarks made in previous chapters about the measurement of vectors $\boldsymbol{z}$ and $\boldsymbol{q}$ (comparability, "condensation", etc.) also holds for $\boldsymbol{w}$.

analogous partial relationships, for simple special cases, could be formulated exactly. (See, for example, formulae (8.13) and (8.17), (B.12), and (B.17) and the models associated with these formulae.)

The explanatory variables of the function only include *productive* (and not the whole) slack. Arbitrary increases of unproductive slack do not decrease the shortage intensity, either. (In the two special cases discussed earlier in sections 8.2 and 8.3, the accumulation of unproductive slack was automatically excluded.

Although we do not know exactly the mathematical form of relationship (8.23), we should also like to express graphically what we have to say. (See fig. 8.6.) The figure, however, does no more than help the reader imagine the nature of the relationship. We hope that this will facilitate understanding of the subsequent argument.

First a few technical remarks about the figure. Each of the three groups of phenomena are described by a vector with many components. Yet the paper is a two-dimensional plane, so that the interrelations between at most three variables can be represented on it. It is only for this reason that each group of phenomena is represented by a single scalar variable. The figure should be considered as one showing the relationship, *ceteris paribus*, of a single shortage indicator, a single productive slack indicator, and a single friction indicator.

The shape of the curves on the figure is arbitrary. Let us recall that the specific friction functions ϕ and ϕ' illustrated in figs. 8.2 and 8.3 were linear, but this was only because of the very simple expository models

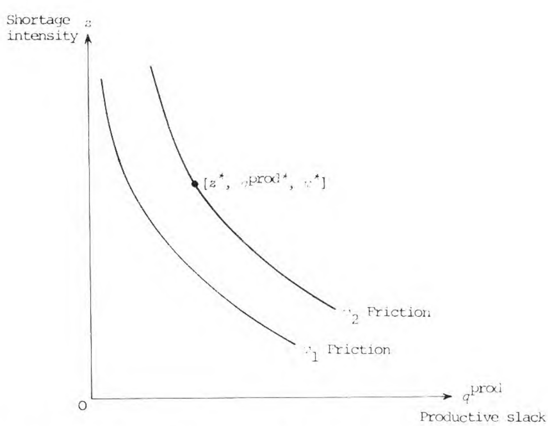


Figure 8.6. General relationship between shortage, productive slack, and friction.

being used. We cannot say anything about the linear or nonlinear nature of the more general friction function.

Let us first consider the origin. The system is in a *Walrasian normal state* if on average over a period there is neither shortage, nor slack,[21] nor friction:

$$z^* = 0; \qquad q^* = 0; \qquad w^* = 0. \tag{8.24}$$

This is not solely a definition but a statement implying a *causal* relation. *The normal state of the system can be Walrasian only if adjustment is perfectly free of friction.*

The origin can be given two contrasting institutional interpretations. The first is the "*perfectly competitive equilibrium*" with completely decentralized decision, with all the conditions, however, that provide for frictionless operation. According to the second interpretation, it is also the position of a *perfectly functioning centralized economy*. Let us assume a system in which all information arrives immediately and with no distortion at the center, where it is processed without delay, leading to instant reactions in the allocation plans, which latter are implemented at once and perfectly. These plans guarantee at all times the full utilization of resources. The origin symbolizes a system functioning with no friction at all, controlled either by a "visible" or an "invisible" hand.

Yet in reality no system exists in which adjustment takes place without any friction at all. There is friction in anonymous market mechanisms as well as in systems of central planning and economic management. It is true that the nature of these frictions varies widely, and each type requires separate study. But none is "perfect" in the above-mentioned sense of $w^* = 0$. That is why no system does or can exist with normal state the origin of the axes shown in fig. 8.6.

In the description of economic adjustment the position $z^* = 0$, $q^* = 0$, and $w^* = 0$ acts as an abstract point of reference.[22] And real systems can be described by the departure of their indicators z, q, and w from the origin. In fig. 8.6 such a point is shown for illustration; this is the *normal*

[21]It is useful to define the Walrasian normal state in such a way that there is neither productive nor unproductive slack in it: $q^* = 0$. The argument of the friction function is, however, only the productive slack; accordingly, only q^{prod} is seen in fig. 8.6. $q^* = 0$ in (8.24) obviously implies the value $q^{\text{prod}*} = 0$, associated with the origin of the figure.

[22]By an ingenious analogy, Siven–Ysander (1973), in their review of *Anti-Equilibrium*, assign to the concept of Walrasian equilibrium a role similar to that played by the absolute freezing point in physics. Although empirically the latter does not exist, it is the theoretical starting point of a measurement scale. The Walrasian state $z^* = 0$, $q^* = 0$, and $w^* = 0$ has a similar role in the above train of thought.

state of some *real system*, permanently at some distance from the position of the Walrasian normal state, namely the origin. The instantaneous states of the real system in question fluctuate around its normal state $z^* > \boldsymbol{0}$, $q^{\text{prod}*} > \boldsymbol{0}$, and $w^* > \boldsymbol{0}$.

8.8. Separation: Friction versus "siphoning off" of slack

The abstract train of thought presented in sections 8.2–8.7 is to help in understanding the functioning of *real* economic systems and, primarily in the analysis of our main subject, *shortage*. Shortage is explained by several intricately interdependent factors. The apparatus of the present chapter enables us to divide the set of explanatory factors into *two*.

Let us consider our customary example: screw manufacturing. We assume that several screw factories exist side by side; screws are sold partly by the factories themselves, and partly they reach the hundreds of consumer firms through the mediation of commercial firms. There is no "general" shortage of screws, but there can be "partial" shortage: the buyer may not be able to obtain either this or that particular type of screw when and where he tries to get it.

At meetings of economic managers as well as in the Press the following kinds of explanation are given: "Screw factories wrongly predicted needs". Or: "The screw factory N.N. delivers erratically; it suddenly delivers too much, or – over a long period – it delivers too little." Or: "Screws are allocated wrongly: regions A and C get too many, while regions B and D too few of the type of screw in question." Or: "The users make their demands known too late; therefore, producers cannot prepare themselves in time for production". "Inconsiderate planning", "disorganization", "bad co-operation", "lack of foresight", "failure to fulfill contractual obligations", "lack of discipline" – such judgements are often passed.

All the explanations enumerated above belong to the group of phenomena that was summarized by the title of this chapter in the following words: *frictions* of adjustment. While we have been quoting statements frequently heard in economic life, the reader can identify each in turn with the situations outlined in the abstract models of sections 8.2–8.6. Those who explain shortage by these friction phenomena speak the truth–but only part of the truth.

Ceteris paribus, an intensified degree of friction increases the intensity of shortage, while its reduction decreases it. This situation is presented in fig. 8.7 which is of identical structure to fig. 8.6. Suppose we shift down the

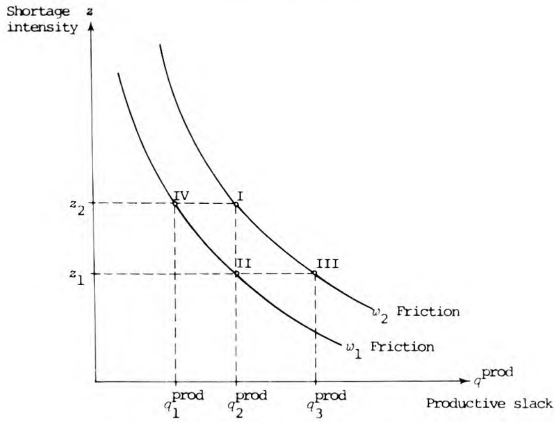


Figure 8.7. Shifts in the normal state.

vertical line above q_2^{prod}. Friction has decreased, that is, the change from w_2 to w_1 has taken place. The screw-sector shifted from the iso-friction curve associated with w_2 to that associated with w_1. Thus, the sector moved from point I to point II, as a result of which the shortage intensity decreased from z_2 to z_1.

In our example, this means that the demand for screws is better assessed, the regional allocation of screws has improved, deliveries became more punctual, and so on. As a result, buyers of screws must wait less and search less, and forced substitution becomes less extensive.

The shift from point I to point II is not the only possibility. Let us consider two further ones.

Let us begin with point III. There has been no change in planning, organization, and discipline in the production and trade of screws, and the prediction of demand has not yet improved. In a word, friction has not changed; the system remains on the iso-friction curve associated with w_2. Yet productive slack has grown. There is a larger output stock in screw factories, larger stocks are held in stock by the trade, consumers keep larger input stocks, and for all of them it is "current" stocks that have grown. Producers have more immediately mobilizable stock: machines, material, and labor, that is, all complementary inputs, are together so that production can be adjusted to the needs of the moment.

As a result, shortage intensity is reduced, since a large part even of changeable, fluctuating, and unforeseen demands can be fulfilled either from stock or through a rapid adjustment of production. That is to say, the change from z_2 to z_1 takes place. *The increase in productive slack has the same effect on shortage as the reduction of friction.*

Let us not omit from our calculation, however, the possibility of a shift in the opposite direction, whereby the system moves from point I to point IV. Now friction has been reduced, with improvements taking place in planning, foresight, and organization. As a result the system moved from the upper iso-friction curve to the lower one. Meanwhile, part of the productive slack was "siphoned off' from production and trade. (That is, the change from q_2^{prod} to q_1^{prod} has taken place.) This may have happened for several reasons, for example the need to fulfill an unusual export order, or the appearance of new consumers on the home market before production grows to fulfill their demand. In the factories there may have been some mobilizable reserve capacity, but now that is also used, and output stocks diminished.

Under such conditions – even if the composition of stocks has improved, and production is better coordinated with consumption – shortage phenomena will appear again and again with the earlier intensity. Even with improved adjustment unexpected demands may arise and disturbances may occur. And, if total stock and total reserve capacity are small everywhere, deviations between actual supply and actual demand may certainly occur, with the associated queuing, waiting, search, and forced substitution. Even if the change from w_2 to w_1 has taken place, it was counterbalanced by the change from q_2^{prod} to q_1^{prod} so that the intensity of shortage has remained at z_2.

This shift from point I to point IV is not just a case invented for the sake of logical completeness. *Although shortage and slack, as well as friction, each have their normal values, it seems that it is the norms of shortage in particular that show the greatest resistance to change. Reduction of friction in some field does not necessarily entail a permanent elimination of shortage in that field.* If in that field the difficulties caused by shortage have become conspicuously rare – queues are shorter, forced substitution is less serious – this may serve as a signal that attention should now be turned to another field. Preference in the allocation of resources must be given to that field where shortage is just as intensive, giving rise to the opposite "voice" – frequent complaints and protests. (We have already mentioned this in Chapter 3, when discussing short-term adjustment, and we will revert to the subject in Chapter 10 in connection with long-term control.)

In the present chapter we divide shortage-inducing factors into two groups, with productive slack as the second group of factors. Needless to say, the volume of productive slack (for given friction) is not the *ultimate* regulator of shortage intensity. On the contrary, it is only the last *link* in a complex causal chain which has a direct effect on the development of shortage phenomena. The causal explanation must make clear the mechanism that constantly "siphons off" productive slack from production and trade. This has already been treated in earlier chapters (recall the concepts of quantity drive, tautness, hoarding tendency, the firm's almost-insatiable demand, etc.), and will be developed further later on (again a few headings, investment hunger, soft budget constraint, etc.). We have referred several times to the various self-reproducing and self-generating vicious circles of shortage. Once again, it is this chain of effects that is "symbolized" in the theoretical apparatus of this chapter (among others in figs. 8.6–8.8) by asking how much productive slack has remained in the system, or how much slack has been "siphoned off" by socioeconomic "suction" processes ultimately explaining shortage.

In the present chapter also the effects of q^{prod} and w on shortage have been distinguished. It must be emphasized, however, that slack and friction are also interactive. *The less the productive slack, the more difficult any kind of adjustment becomes.* Productive slack may serve, as has been mentioned, as a buffer for the localization of various disturbances. If there is an inadequate buffer, every friction will spill over much more. One explanation (not the only one) for persistent inflexibility is the instinctive self-defense of the system against the spill-over of friction.

This connection between slack and friction is represented by the shape of the iso-friction curve: it grows steeper to the left. The smaller the productive slack, with a given degree of organization and adjustment abilities, the more frequent and severe the shortage phenomena that will occur.

With the aid of the preceding apparatus we may consider the significance of the *Hungarian economic management reform of* 1968 from the point of view of the interrelations being discussed here. The elimination of a considerable proportion of bureaucratic restrictions, and the greater independence of firms and nonprofit institutions, diminished the frictions of adjustment. In fig. 8.8 a shift took place in the direction of the arrow pointing downwards, moving the system to the lower iso-friction curve representing less friction.

This change coincided historically with changes taking place in economic growth policy in the 1960s. Production and investment plans became moderate in comparison with those of the 1950s, "quantity drive" was asserted less strongly, plans became less taut. All that involved, too, a lessening of suction, that is the siphoning off of slack from production and

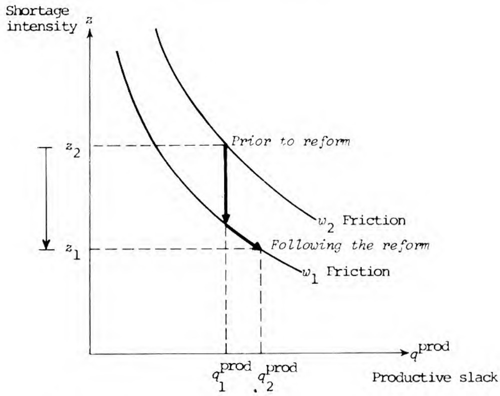


Figure 8.8. Impact of the reform of economic management or friction and shortage.

trade. This is represented in our figure by a shift to the right along a given iso-friction curve.

The two different changes have to be added or multiplied,[23] and from their joint effect, shortage intensity has lessened. Shortage has not disappeared, but in numerous fields shortage phenomena have become rarer as well as less severe. Although it is too early to be sure, it seems as if not only the present actual state has moved, but also as if the *norms* of shortage have shifted. Today – particularly on the consumer market of which we shall talk in Part II of the book – "shortage norms" are different from those of ten or fifteen years ago.

Of course, this is all a hypothesis which needs careful testing. It not only requires more detailed studies, but also a longer historical perspective. Many of the elements of the suction-reproducing mechanism continue to exist, as mentioned several times in the book. Decreasing shortage may be counterbalanced by increasing suction. The case symbolized by the shift from I to IV in fig. 8.7 cannot be excluded from future possibilities.

Finally, one last remark, not on the Hungarian reform but on the general topic of the chapter. Friction is present in *every* system, while the phenomenon of suction (and with it, the repeated tendency of siphoning off the productive slack) is *system-specific*. It is important to make this distinction already for that reason. Of course, friction takes particular

[23]The two terms indicating mathematical operations are used "allegorically". As we explained earlier, we do not yet know well enough the exact nature of general friction functions.

forms in each concrete system, and there is no real economy without friction. Even where stores are packed with goods, and where ample rapidly mobilizable reserve capacities do exist, there may be a buyer who happens not to find some product when and where he wants to buy it. This book seeks to identify the mechanisms that amplify this everywhere existing friction-caused shortage and reproduce frequent, intensive shortage phenomena rather than sporadic, modest ones.

8.9. On the "short side rule"

All that has been said about friction, shortage, and slack has further *theoretical implications*. At this point I wish to reconsider one of the most important categories of microeconomics, the concept of "excess demand", in the light of the discussion in Chapters 7 and 8.

The concept of "excess demand" for product i is usually defined in the following way:

$$\underset{\boxed{\substack{\text{excess}\\ \text{demand}}}}{e_i} = \underset{\boxed{\text{demand}}}{d_i} - \underset{\boxed{\text{supply}}}{s_i}. \tag{8.25}$$

Excess demand is positive if demand exceeds supply, and is negative if supply exceeds demand. In other words, excess supply can be described as negative excess demand.

However simple this formula may be, the mere fact that the operation of *subtraction* is applied assumes the fulfilment of certain conditions. Subtraction is justified if the following rule is asserted:

$$\underset{\boxed{\text{purchase}}}{y_i} \leq \underset{\boxed{\text{demand}}}{d_i}, \tag{8.26}$$

that is to say, the buyer does not purchase anything for which he has no demand. In an even more general form:

$$y_i = \begin{cases} d_i, & \text{if } d_i \leqq s_i \quad \boxed{\text{excess supply (or equality)}} \\ s_i, & \text{if } d_i \geqq s_i \quad \boxed{\text{excess demand (or equality)}}. \end{cases} \tag{8.27}$$

This rule is also called the "short side rule", and it has already been mentioned in another context.

It is basically an empirical question whether the "short side rule" holds. If it does (or does so to a good approximation), it is justified to deduct supply from demand to determine the size of excess demand. If, empirically, it does not hold, this procedure is not justified.

We attempt to answer the question at two different levels. The first is the *submicro-level.*[24] We visit a *single seller* and observe the *instant* at which a *certain buyer* looks for a *particular product*. "I want 100 tons of steel of quality 10." If at the moment there are 250 tons of steel of quality 10 in the seller's store, the buyer will receive his 100 tons. The final state of the instant is -150 tons of excess demand (this negative number indicates the quantity remaining in the seller's store after the transaction). If, however, there were only 50 tons in the seller's store, the buyer will take this quantity, and the final state will be $+50$ tons of excess demand (and the positive number now indicates unsatisfied demand).

Experience also justifies the claim that *at the submicro-level the "short side rule" is generally asserted*. Even though exceptions occur (e.g. in the case of indivisible complementary actions),[25] the rule generally prevails.

The description of a partial market according to the following rules is referred to as the *micro-level* (above the submicro-level).

(a) Not a single instant, but a finite period of time is observed.

(b) Instead of a single buyer meeting a single seller, all buyers in the partial market, meeting all sellers, are observed.

(c) Not just the first act of the shopping process (whether or not purchase in accordance with initial demand takes place), but possibly also the acts following the first one, namely forced substitution, search, and so on, are described.

In the case of a micro-level description empirical observations suggest that *the short side rule asserts itself only exceptionally*; *more often it does not*. The buyer may buy more of the substitute than his initial demand, when he applies forced substitution. (Violation of condition (8.26).) As a consequence of friction phenomena shortage *as well as* slack may be present on the same partial market within the same period (at different stores). (Violation of condition (8.27).) And, if all that is so, the empirical back-

[24] We now examine the fulfilment of buying intentions on the market, the meeting place of buyer and seller. Previously, in section 2.10, we analyzed an analogous problem in connection with the producer's intentions regarding input use.

[25] "If I do not receive 100 tons, I prefer not to take 50 tons, because it is difficult to arrange transport."

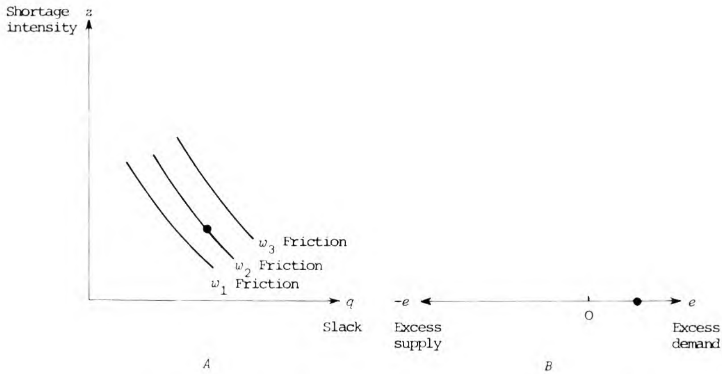


Figure 8.9. Description of the state of the market in multidimensional or one-dimensional space.

ground to the deduction formula, "demand minus supply equals excess demand", together with its logical clarity, becomes doubtful.

I believe that these doubts touch upon the *foundations of standard microeconomics*[26] and in this respect they do not concern only the analysis of a resourse-constrained economy. There is friction in the adjustment of all economic systems, therefore in all of them forced substitution as well as shortage and slack appear. It is true that the frequency, intensity and distribution of these phenomena are widely different within each system, but their mere existence is general. That is why I would consider the foregoing criticism of the theoretical foundations of microeconomics as a general criticism pointing beyond the problems of the shortage economy.

The category of "excess demand" defined in (8.25) describes a group of phenomena in *one-dimensional* space which can only be described in *multidimensional* space.[27] Let us examine fig. 8.9. Part A is a simple repetition of fig. 8.6, and all that was said about the latter holds for this,

[26]As a matter of fact, it is not quite clear at what level the theorems of standard microeconomics should be understood. For example, the description in Debreu (1959) may be considered as strictly at a submicro-level, since products are "labelled" separately by date and area. The conventional interpretation of microeconomics implies, however – at least tacitly – a micro-level description.

[27]To avoid misunderstandings it is *not* the model of the partial market trading a single product that is opposed to models with many products. It is not *this* difference we are talking about when we mention the number of dimensions. A model of a partial market is compared with a model of another partial market. Both models deal with the allocation process affecting the trade in a group of products, and describe the process in one dimension in one case and in many dimensions in the other case.

too: it is the two-dimensional character of paper that forces us to show only a scalar variable for each of the vectors $\boldsymbol{z}$, $\boldsymbol{q}$, and $\boldsymbol{w}$, all three consisting of many components. The point shown on the figure is the state of the market represented in the multidimensional space.

Part B of the figure illustrates the approach of standard microeconomics. With this apparatus the state of the market is described by a point on the *excess demand line*: it is either to the right, on the positive half-line, or to the left, on the negative half-line. As opposed to the *multi*dimensional approach of part A, that of part B is *one*-dimensional.

Finally, one more remark on the *Clower–Barro–Grossman school*. It directs great attention to states of the market which deviate from the Walrasian equilibrium. Yet it does not break away from the "short side rule"; moreover, it even stresses the rule as one of the cornerstones of its theory.[28] In the case of shortage, the buyer is not satisfied because his demand meets the physical supply constraint. Yet in reality, although this may often be the case, dissatisfaction may also be the consequence of the fact that seller and buyer, that is the initial demand and supply, fail to find each other. The microeconomic approach of the Clower–Barro–Grossman school basically describes states deviating from Walrasian equilibrium using the one-dimensional approach of fig. 8.9B and does not enter the interior of the multidimensional space spanned by the vectors $\boldsymbol{z}$, $\boldsymbol{q}$, and $\boldsymbol{w}$.

[28]See, for example, Clower (1965) and Benassy (1974, 1975, 1977). I must make a self-critical remark here. In my book *Anti-Equilibrium* I committed myself the same inexactitude as I noted above. (See Kornai, 1971a, b, ch. 19.) I ignored the possibility that, as a consequence of frictions in adjustment, shortage and slack may occur simultaneously. The intentions of both buyer and seller may remain unfulfilled at the same time.

CHAPTER 9

The investor: Institutional framework

9.1. Introduction

The approach of the foregoing chapters has already been dynamic: production, purchase, selling and their mutual adjustment have been described as processes over time. Yet the question has been left open up to now as to whether production and trade repeat themselves in identical volumes, or are growing. In Chapters 9 and 10 our analysis will move forward and discuss investment. We do not aim at completeness by any means; only one or two aspects of the vast complex of possible questions will be dealt with. We will focus on issues related to the main subjects of this book, that is on shortage and on "quantity" adjustment.[1]

Chapters 9 and 10 belong closely together. In Chapter 10 we shall discuss the signals influencing investment allocation among specific projects, among firms, nonprofit institutions, and sectors. The long-term adjustment of production to the needs and requirements of users will also be discussed there. Later, in Chapter 13, we shall once more revert to the question of selection criteria. In the present chapter we put aside this set of issues: we do not ask what are the *contents* of the investment decision. That is because firstly we wish to look at the *institutional framework* of the

[1]In narrowing down the issues to be discussed I was influenced by another point of view. Several problems of the growth of a socialist economy were treated in my book *Rush versus Harmonic Growth* (Kornai, 1972a, b). I do not wish to repeat in this work what was said there. The ideas of my earlier book and the present Chapters 9, 10, and 12 are closely linked.

A large number of studies have been written on the investments of a socialist economy. Outstanding among them is Bauer (1977) not only because of its vast empirical material but also its thorough theoretical analysis. All through Chapter 9 and particularly in the last section treating cycles I largely relied on Bauer's book, as well as on one of his earlier preliminary studies, Bauer (1975b).

I used the rich literature on investment problems, especially the studies by Soós (1974, 1975a, b), Deák (1975, 1978a, b), Huszár – Mandel (1975), and also the book by Faluvégi (1977).

investment decision, and the *motivation* of the decision-makers. The common title of the two chapters ("the investor") indicates that it is the decision-makers' behavior that is at the center of the examination.[2] Although the changes in the real sphere (for example, the relative sizes of different sectors, technical development, etc.) are important, at present we are only analyzing the *control* sphere.

So far we have talked exclusively about the behavior of *firms*. We now extend our analysis to *nonprofit institutions*. The institutions which come into this category are those which do not cover their expenses from selling products or services for money, but receive other resources such as a state subsidy, donations, and so on. For example, in Hungary every school, university, hospital, and most research institutes are in this category. We also consider as nonprofit institutions – at least from the point of view of the subject treated in Chapters 9 and 10 – the organizationally separated units of central and local administration, namely each organization that has a separate budget and financial accounting. (For example the county Board of Monuments Protection or the city fire service.) We do not think that firms and nonprofit institutions behave in an identical manner, but, in the case of investment, their behavior is similar.

Most of the chapter is concerned with the *traditional* economic management system of the socialist economy. We briefly point out in sections 9.8 and 9.9 how much the situation has changed in Hungary since the reform.

Even concentrating on the traditional system, this is not homogeneous, or unchanged over time. In different socialist countries, and within each country at various historical periods, specific organizational patterns differ from each other, as do the delimitation of spheres of authority, rules of material incentives, and so on. It is not one specific system that we are going to describe, but a somewhat "stylized" general model of traditional control.

In the traditional economic management system the firm and the nonprofit institution cannot make independent investment decisions,[3] that right being reserved for superior authorities. The exact division of competence varies; it depends on the sectoral character as well as on the size and importance of the particular investment project. In the case of the largest and most important investments the decision is made by the government; in the case of other investments by the Minister or by

[2] I use the word "investor" as a collective term to name not only the firm or non-profit institution directly executing the investment, but all individuals or bodies making decisions about investments.

[3] There may be exceptions, such as part of machine replacement financed from depreciation funds.

the director of the sector. One part of decision-making is connected with the elaboration of yearly and five-year plans. The other part – independently of the schedule for preparing the economy-wide plan – concerns individual investment proposals.

Since we do not wish to go into details, we can give a satisfactory account of the vertical linkages of the control process, if – as a simplifying assumption – we describe investment decisions in the following organizational framework.

We have in mind a group of several firms or of several nonprofit institutions. They are the *claimants*, competing with each other for investment resources at the disposal of their common superior authority, called here the *allocator*. The question is left open who is in fact the allocator, what organ or authority, and whether the decision-maker is an individual or a collective body. In our description the allocator "personifies" the one who makes investment decisions.[4]

Although there is a super- and subordination relation between them, it is not only the allocator who influences claimants but also claimants influence the allocator. Decision is preceded by discussions, repeated exchange of information, arguments, and bargaining. Let us first examine the *claimants*' behavior.

9.2. The claimants' behavior: Expansion drive and investment hunger

Let us begin with a general observation.

In a socialist economy there is no firm or nonprofit institution which does not want to invest. There is no saturation. *Investment hunger* is permanent. If an investment project just finished has appeased hunger momentarily, it will shortly reappear, and more intensively than ever.

I could not say where the Hungarian expression "investment hunger" comes from. Has it been taken up by theoretical studies from practical discussions, or was it the other way round? The expression is a graphic one. It shows very well that we are talking about something resembling the elementary needs of living organisms which break through irresistibly and constantly.

Marx mentions the capitalist's "animal instinct" which drives him towards accumulation in the hunt after profit. But what is it that drives the

[4] It will be explained in section 9.7 that the direct superior authority of the firm or of the nonprofit institution is not sovereign itself: the lower-level allocator is subordinated to the medium-level and high-level allocator.

socialist economic leader – not interested in profit – toward investment and real capital accumulation? In Part II of the book we shall look at the effect of profit-seeking in a socialist economy. For the time being we neglect this aspect and think of the traditional case in which the material and moral interests of the firm are not linked to profit. (The nonprofit institution comes into this same category by definition.)

We have to repeat ourselves to a certain extent as regards section 3.6. What was said there about firm managers' motivation in relation to *current production* also holds for *investment*. The most important motive is that the leader – whether he is the responsible manager of a workshop, a producer firm, a hospital, or a school – *identifies with his job*. He is convinced that the activity of the unit under his charge is important. Therefore it has to grow. He is pressed by the internal problems of the unit and believes that they could be solved, at least partially, by investment. These machines are out of date – it is time to replace them. The foundry is unable to supply enough casting to the metal cutting workshop – it ought to be enlarged. The surgery needs new equipment. Projectors ought to be acquired in order to modernize education.

One's own unit always seems outdated and poor, in comparison to a similar unit which is better equipped and more modern, either in our country or in a foreign country. The leader feels *professional envy* in the good sense of the word. He wishes to augment his *professional prestige*: he would like to show off the new machine, the new workshop, or the new building. Investment is needed so that this desire can be realized.

These feelings are intensified by *shortages*. Thus, one of the foregoing examples: the disproportion between the foundry and the metal cutting workshop pointed to shortage *inside* the unit put under the leader's charge. But shortage also makes itself felt *outside*. There is queuing for the firm's products, buyers impatiently demanding more and more. Investment is needed so that those in the queue can get more. The same is felt by the hospital director who cannot admit all the patients waiting for hospital care, and by the University Rector who cannot admit all young people wanting to study. This feeling is further amplified by the fact that there are quite a number of firms or nonprofit institutions which have been officially declared solely responsible for supplying the economy or the population with a certain product or service.

So far only noble and unselfish motives have been mentioned. They may be accompanied by other less noble, but quite human and understandable motives. The leader's *power*, social prestige, and consequently his own importance grows together with the growth of the firm or nonprofit institution. Many feel that it is greater to be the director of 10,000 than of 5,000. A greater power may be concomitant – depending on the actual system of rewards – with *larger material acknowledgement*: higher salary, bonuses, and privileges. It is not worth shutting our eyes to the fact that, consciously or unconsciously, such considerations may also play a role in making the leader of the nonprofit institution or the firm fight for the growth of his unit. These are, however, motives of secondary importance which do not apply to everybody. If, for example, someone has become Rector of a large university, or is responsible for protection of all monuments, or is charged with flood-prevention for the whole country, neither his payment, nor his authority or power will grow any further if he can obtain 20 percent more investment for his own field. And yet he will fight like a lion for such investment.

When wanting to invest, the leader of the firm or of the nonprofit institution knows his men are with him: it is not against their will but in most cases with their full agreement or perhaps even expressly with their encouragement that he is acting. In fact, one of the usual criteria of a "good leader" for subordinates is someone who can secure a lot of investment for the concern under his charge.

To summarize we shall refer to the joint effect of the motivations listed above as expansion drive. It is a form of behavior which was preconditioned by social relations, but became so deeply rooted in the thinking and acting routine of leaders of firms and nonprofit institutions in the socialist economy that it has become a "natural instinct".[5] One *must* grow. Expansion drive joins quantity drive and hoarding tendency in creating the state of suction and the almost-insatiable demand in the production sphere. (Quantity drive and hoarding tendency were discussed in sections 3.2 and 5.6 in relation to current production.) Expansion drive is even more important than quantity drive and hoarding tendency since its effect on the operation of the system is even stronger. *It is expansion drive that explains insatiable investment hunger.*

Expansion drive is found at all levels of the economic hierarchy: from the head of a team consisting of just a few workers up to the Minister in control of a hundred thousand or a million. When it comes to the distribution of investment resources, each fights for more investment for

[5]The mechanism and motivations of expansion drive are illustratively shown in the case study by Laki (1978b).

our team, *our* firm, *our* ministry. Expansion drive is revealed equally in discussions among shop-foremen within a firm, and at government meetings of ministers responsible for various large economic sectors.

Expansion drive would push investment demand to infinity. The question arises whether any factors exist which would limit demand. It is not our aim to point out that infinite demand cannot be satisfied by limited physical resources nor that some claims would be rejected by the allocator in the course of the approval procedure for investment projects. The question therefore arises whether there are any factors which would encourage claimants to limit their claims *voluntarily*.

Before giving a positive answer we must make a negative assertion. The *demand for investment is not limited by fear of loss or failure*. This is a logical consequence of the fact that investment comes as a *gift* for the claimant. In the case of the nonprofit institution this is clear and obvious. In the case of the firm it may be covered by "money illusion". Depending on the actual system of accounting a "credit" may be given for investment which has to be repaid. In Part II of the book this will be analyzed in detail. Despite this arrangement the firm considers investment as a gift. It feels sure that repayment of money received for investment purposes can never cause much worry. There are, of course, more successful and less successful projects. It may be that an investment decision or its implementation is criticized afterwards. *Yet, failure, in the true sense of the word, is impossible*. A genuine failure would arise if the additional production resulting from the investment was impossible to sell. Or, if costs were so high that the investment does not pay off, the firm would go bankrupt. *This kind* of failure cannot happen. This is guaranteed, on the one hand, by shortages since every product is saleable. If the buyer does not buy it voluntarily, he will, sooner or later, buy it in forced substitution. On the other hand, a possible financial loss suffered in home or foreign sales can always be compensated – as we shall explain in a subsequent chapter – by state subsidies, price adjustment, or other means. The budget constraint for expenditure earmarked for an investment project is soft. We shall discuss this further in Chapter 13. The result is that every investment will automatically prove justified.

Since every firm and nonprofit institution is affected by expansion drive, with no failure threatening to keep anybody back from expansion, nobody renounces investment *voluntarily*. Managers in the socialist economy are so used to this situation that they cannot even imagine that anything different is possible. And yet it is exactly there that one of the most important differences lies in comparison with the mechanism of the capitalist system.

To invest or not to invest – that is one of the gravest dilemmas of the capitalist firm. It may shy away *voluntarily* from the thought of investment if it seems too risky, even if it has got momentarily the required financial cover. This kind of reluctance is unknown in the traditional management system of the socialist economy.

There are other reasons why demand is not infinite. The initiator of the investment knows perfectly well that there is a tacit upper limit above which it would not be advisable for him to make a claim. Obtaining investment is a complicated campaign which necessitates clever maneuvering. The claimant is motivated to self-restraint by "tactical considerations". No absurd volume of investment must be asked for, and no new claim must be put in the day after finishing the previous large project. That is why it can be said (and it is important to know also from the theoretical point of view) that, although investment claims would be pushed towards infinity by the motives explained above, claims are *limited* because of tactical self-restraint. The sizes of claims are observable and measurable.

Although claims are limited, they are always well in excess of available resources. This situation is expressed as follows: *demand for investment resources is almost insatiable*. (In Chapter 5 we talked about the similar almost-insatiable demand for inputs for current production.) *This almost-insatiable demand naturally affects the social relations between claimants and allocator*.

9.3. The allocator's behavior: The tautness of the investment plan

In describing the allocator's behavior let us start from the fact that his own superior authorities have prescribed investment quotas for him to disperse. The quotas specify how much in total can be spent by the sector belonging to the allocator within a given period (one year or five years) and how much can be spent within the aggregate quota on the main types of expenditure: construction, machines made by domestic industry, machines imported from socialist or from capitalist countries, and so on. We postpone looking at how rigid or expandable the quota is. And we also examine later how the quotas, aggregated at current or fixed prices, are related to the physical constraints of the resources actually available. For the time being it is assumed that the quota is given for the allocator.

It occurs to the allocator that it might be as well to leave a small reserve. Perhaps the quota should not be distributed to the last penny among claimants, since subsequent claims may still appear. What size should this reserve be? It seems unimaginable that, for example, 20 to 30 percent of the quota should be left undistributed. The expansion drive is no less strong in the allocator than it is in his subordinate claimants. If the allocator is likely to go far, he will probably start with the intention of leaving between 3 and 5 percent undistributed.

Selection starts. (This will be discussed in more detail in Chapter 10.) Huge pressure is put on the allocator. Each claimant affirms repeatedly the indispensability of his investment. The quota is inevitably filled up with permitted individual investment projects. If the allocator was very firm, he will have kept the 2–5 percent reserve. But if he was more lenient, he will have gone to the upper limits of the quota.

Seemingly everything is in order; investment resources are in exact *ex ante* balance: all resources are totally utilized. Yet further developments are not so ideal.

Nearly all claimants underestimate expected costs. Psychoanalysis might reveal whether this distortion is conscious and deliberate, or if the claimant is driven by his "subconscious". Undoubtedly he has an interest in underestimating costs since the chances of acceptance of the project are greater if expected costs are relatively smaller. The discrepancy between estimated costs and actual ones is not of the order of 2 to 5 percent, that is the order of the resources reserved by the comparatively cautious allocator. The average may be about 20 to 30 percent, and a discrepancy of 50 percent is not infrequent. A part of this discrepancy is connected with rises in prices, while another part reflects *actual physical inputs* exceeding estimates.

The regular and substantial surpassing of the planned level of inputs would in itself fully explain a permanent shortage on the market for investment goods. And yet this is supplemented by another phenomenon. Projects *outside the plan* are always added to the originally approved investments. Several factors may play a role in this, a few of which will be mentioned.

(a) Unforeseen troubles and difficulties arise, such as floods or other natural disasters, a sudden deterioration in the foreign trade situation, unexpected problems in one of the domestic sectors, etc. Adjustment requires – among other things – quick investment.

(b) A number of projects must be carried out outside the plan which, if there had been better foresight, could have been included in the plan from the outset. The construction of the new factory is started and it turns out that auxiliary works as well as joint infrastructural and social projects are needed. The manufacture of an important new final product is started through a large investment, and again it is observed only in the course of development that the "hinterland" industry must be built up to provide semifinished products or parts for the new final products. The question arises whether the planning error was due to deliberate distortion or to forgetfulness caused by the "subconscious". From the "tactical" point of view it is obviously easier to get permission for a relatively smaller investment (for the construction of the basic factory only; for the organization of manufacturing of the final product only).

(c) Unforeseen favorable opportunities present themselves. For example, we could break into a foreign market which has just been opened, or an invention is made. Prompt investment is needed to grasp the favorable opportunity.

(d) Political leaders, members of Parliament or of the local council make pledges to a group of people, promising the construction of a new hospital, or a new road. After that they put pressure on the economic apparatus to fulfill the pledge.

All four types of event would require additional investment over the quotas already exhausted. A Hungarian proverb says: "There is no haycart that won't carry just one more piece of hay." And, in fact, since a single additional investment is such a small item in itself, why could it not be squeezed in among the projects approved? Some of them are refused, but others are allowed.

The problems described above have a characteristic *time* dimension, felt differently by the claimant and the allocator.

From the *claimant's point of view* investment is a long campaign with many battles. But the whole campaign has only one life-and-death battle and that is at the beginning, since approval must be obtained for *starting* the investment. Once started, it will end in some way and at some time. That is exactly why it is possible to underestimate, without much hesitation, expected costs, and to forget about complementary investments. If costs are higher, or if investments above the plan are necessary, money will surely be raised in one way or another. Perhaps the claimant will be blamed for erroneous calculations, perhaps work will slow down for a while to wait for financial cover, but an investment project that has been started will not be stopped for good.

All this means *from the allocator's point of view* that before each plan period a considerable part of the investment quota set aside for the coming period is taken up by the continuation and finishing of projects that were approved and begun earlier. The more frequently the above two phenomena occur (surpassing of planned inputs and subsequent squeezing of additional investments into the set of approved projects), the more quotas disperse among the numerous projects started. Because of this investments will be even more drawn out, and the final result will be a larger *committed* share in the investment quota of the following plan period. In other words, the narrower will be the allocator's freedom in assigning new investment projects.

All this makes long-term adjustment more rigid. The economy will only have very limited possibilities of adjusting itself to permanently changed conditions by means of investment. Of the above-mentioned factors, (a)–(d), the first three at least would require fast investment action. Claimants refused usually complain about the plan bureaucracy. They think their claim could be satisfied if planning were more "flexible". This may be a factor but it is not the true explanation. *The control mechanism of the investment process which has just been described necessarily leads to a taut investment plan* (using the concept of "tautness" as explained in Chapter 3). And, if the plan is taut, the planner is *unable* to make flexible adjustments. This means that there are no easily mobilizable slacks, flexibly used spare investment resources to meet unexpected targets. Actions approved and begun earlier have used up almost all investment resources.

The control mechanism of the investment process described above is covered by a "money veil". *Seemingly* quotas established in financial terms are distributed, investment budgets are approved, credits are granted, and so on. *In fact* it is *permission to begin* the actual physical investment project that is granted. *The mechanism necessarily leads to more projects being started than can be performed smoothly and comfortably* (*leaving some reserves*) *within the actual physical resource constraints. Life then squeezes investment projects, amid friction and by delaying each action, into the feasible set allowed by the physical constraints.*

9.4. Investment as a production process, and the investment goods' market

We have reached the point in our explanation where the allocator has approved investment. With official approval of his claim, the claimant now sets about realizing it. And from now on he assumes a new role. So far he has been faced, as the claimant of an investment permit, with the allocator, that is the superior administrative authority. But now he becomes the *executor of the investment*. In some cases the organization of the investment project is done by the firm or nonprofit institution that will ultimately operate the new fixed capital under construction. In other cases a separate "investment firm" is established and committed to carry out the action.

The investment executor must acquire inputs. He becomes a *buyer* who buys investment goods (machines, building activity, installation services, etc.) from the *seller*. For some of the goods the seller is a *commercial* firm (for example, it acquires imported machines from a foreign trade company). Other products and services are, on the other hand, sold to the buyer by domestic *producer* firms such as engineering factories and construction firms. In the traditional economic management system most investment goods are centrally rationed materials. The investment executor appears again as *claimant* and a superior authority plays the role of *allocator* – but now this relationship of "claimant–allocator" has shifted to the plane of "current production".

Thereby we have returned to all the phenomena discussed in Chapters 2–8. Events of investment can be, as it were, "transposed" into the framework used in earlier chapters. That is why, at this point, it is not necessary to go into details.

In a certain sense there is nothing "special"[6] in the execution of an investment. It is no different from a type of current production in which there is vertical multilevel control of production as well as the horizontal relationships of seller and buyer, that is the customary partial market. That is why the following statement is justified: *if the economic system is resource-constrained in current production, so is the investment sphere. Accordingly, all the accompanying phenomena, regularities, and control mechanisms of a resource-constrained system appear here in the usual form.*

9.5. Investment tension

We are now ready to *summarize* our initial propositions, although with a few repetitions.

Hungarian economic managers and theoretical economists both use the expression, *investment tension*. This is a concise expression for a complicated group of phenomena that we have tried to break down into three interactive components.

[6]Attention is drawn to the words: "*in a certain sense*". In *another* sense, obviously, investment activity is indeed a very special sphere, essentially different from all others. This follows from the special role it plays in the expansion of the firm or nonprofit institution, and in the growth of the entire national economy. That is why a separate apparatus deals with investment at every level of economic management and why the investment plan has a separate chapter in economy-wide plans. And that is why we also discuss the special problems of investment in this chapter and Chapter 10.

(1) *In the course of the official approval of investment projects the total of claims always surpasses prescribed investment quotas. There is a tension between claim and quota.*

(2) *Many investment projects approved cannot be carried out with the planned input–output combination and by the planned schedule. The investment plan is taut.*

(3) *The initial demands of firms and nonprofit institutions whose investment projects have been officially approved cannot be fully satisfied from the physical supplies from firms producing and selling investment goods and services. There is a tension between initial demand and actually available real resources.*

Investment tension leads, on the one hand, to a lack of mobilizable slack that could be used to meet unplanned investment demand. There is no "free capital" for unexpected investment. On the other hand, in a state of investment tension the production of investment goods often hits bottlenecks which result in the absence of complementary inputs in the formation of nonmobilizable, unproductive slack.

After clarification of the concept of "investment tension" we can make the following claim, partly on the basis of the foregoing and partly in anticipation of a few ideas in the next part of the chapter.

In the socialist economy *investment tension is continuously reproduced,* although the particular institutional conditions and the central economic policy influence the degree of tension. If, throughout a long historical period, neither institutional conditions nor central economic policy changed essentially, then within that period the *normal* degree of investment tension, as an attribute of *the normal state in the investment sphere,* establishes itself as a permanent tendency.

The normal state of the investment sphere is a vectoral category. It can be described by a suitably selected ensemble of shortage ($\mathbf{z}$), slack ($\mathbf{s}$), and friction ($\mathbf{w}$) indicators. Thus, for example, in the relationship "claimant versus allocator" the ratio of refused demands within total demand is observable; so too is the ratio of building projects not included in the detailed current building industrial plan (but figuring in the original investment programs) within the total program, and so on. These are the characteristic indicators of "vertical shortage". In the relationship of "seller versus buyer of investment goods" the proportion of refused orders to the total number of orders is observable, as well as waiting times, forced substitution (among other things in the form of changes in the original technological plan), the frequency and distribution of bottlenecks in the firm producing investment goods or supplying services. These are the indicators of "horizontal shortage".[7] We do not cite an example for measurement of slack and friction. As follows from preceding chapters, this also seems to be feasible in the investment sphere. It is the intertemporal averages of all these indicators, that is the normal values of the vectors (z^*, q^*, and w^*) that express the normal degree of investment tension in the historical period under discussion.

Shortages present in current production and trade which we examined in Chapters 2–8, and investment tension which has been the subject of the present chapter, are in close interaction: they form a special "vicious circle". Awareness of shortage is one of the main motives for expansion drive and the associated investment hunger. Shortage signals play an important role in the selection of investments. Thus, shortage generates investment tension.

At the same time investment tension is one of the main causes of general shortage. Since investment hunger is insatiable, it creates an almost-insatiable demand. This extends as far as the resource constraints on investment activities, and even goes beyond them. There are no sharp borderlines. Labor, materials, and foreign exchange are shared by investment, household and governmental consumption, exports, and so on. The stronger the investment tension, the more it is felt that investment demand tries to draw resources away from other fields of utilization thus amplifying general shortages.

Economic policy and planning may, in the long run, affect the normal degree of investment tension and, in the short run, the instantaneous deviation from the normal degree. Yet the *basic* phenomenon itself – the existence and continuous reproduction of investment tension – can*not* be explained by planning mistakes,[8] but necessarily appears with the given institutional conditions.

If there were no other factor making its influence felt in this direction, investment hunger and its result, investment tension, would be enough to transform a system into shortage economy. It was emphasized at the end of the preceding section that if the system is resource-constrained in the sphere of current production, it is the same also in the investment sphere. Inversion of the statement holds as well and expresses an even more important interaction, namely *if the system is resource-constrained in the investment sphere, it must also be the same in the sphere of current production*.

[7] The distinction between vertical and horizontal shortage was introduced in section 5.7.

[8] In the literature we repeatedly meet the view that investment tension is caused by mistakes in planning. See, for example, the article by Fonál (1973).

In Keynesian macroeconomics and in the economic policy of capitalist states (if conceived in the Keynesian spirit) the main worry is how investment is to be stimulated when entrepreneurs do not have enough willingness to invest, and how the state can be directed toward more investment activity. Mainly because investment is not sufficient, there is not enough aggregate demand. This problem is unknown in the socialist economy. Investment intentions need never be stimulated since there is a permanent self-stimulation.

Expansion drive and insatiable investment hunger are the main reasons why the productive forces of the socialist economy grow, either slower or faster, but *incessantly*. (We disregard here war damages or the consequences of natural disasters.) In a socialist economy growth is forceful even at times when the production of a capitalist economy declines because of demand constraints. This irresistibility of growth is one of the most important achievements of the socialist economy. It must be added, however, that the *same* deeply effective momentum which renders growth unending also leads to the permanent reproduction of shortage.

9.6. Multilevel control of the investment sphere

Our remarks concerning investment tension interrupted the description of control. Before progressing further, we shall review the structure of processes controlling the investment sphere in fig. 9.1. This promises to be all the more useful since we are concerned with a complicated control acting in more than one dimension. It is worth clarifying what we have accomplished so far, and what it is that remains to be explained.

We can see two kinds of *vertical* arrangement. Let us look first at the right-hand one which represents the distribution of investment quotas and the approval of investment projects. The lowest, fourth level is the firm or the nonprofit institution in which investment takes place and which will operate the real capital formed by the investment project. Before starting the project the firm or nonprofit institution A approaches, as claimant, the superior authority, that is its superordinate allocator placed at the third level in the vertical arrangement. It is not the only one presenting itself, but does so together with units B, C, and D standing at an identical horizontal level. Here it is A, B, C, and D that *compete* for their common allocator's investment quotas. (This was discussed in sections 9.2 and 9.3.)

Let us assume that the claims made by A and B have been approved, while those of C and D have been refused. From that point a new *horizontal* relation appears. Firms A and B are faced, as buyers, with firms

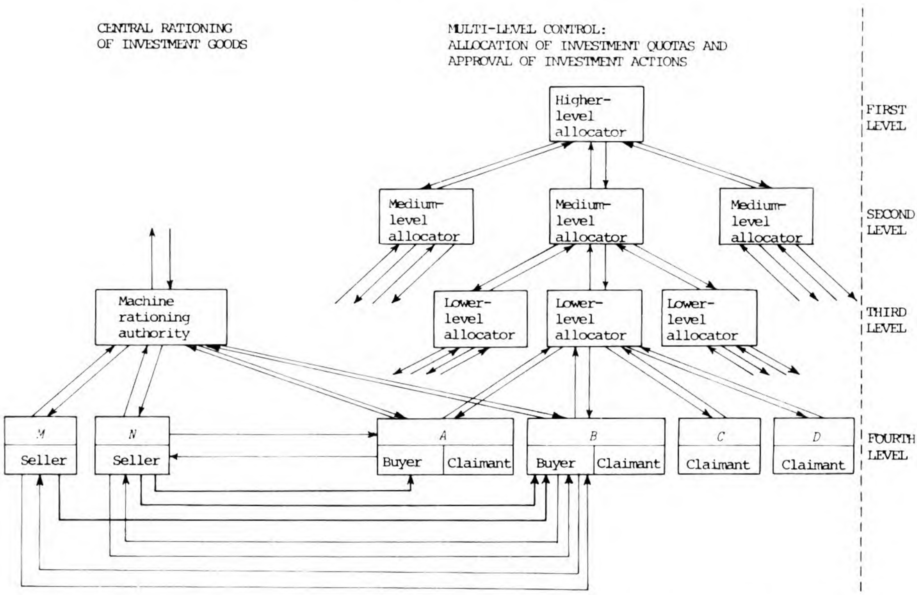


Figure 9.1. Vertical and horizontal linkages in the traditional control of investment processes.

M and N (for example two engineering factories) as sellers. This is a market relation. In this market relation A and B continue to compete with each other ("buyer's competition") for the benevolence of sellers M and N serving both of them. In the figure thin arrows mark exchanges of information and thick arrows transactions of actual physical products or services. The latter pass from M and N to A and B. The units C and D are outside *this* horizontal relationship since their investment claim was not approved.

The horizontal relationship of the producer and user or seller and buyer of investment goods is greatly influenced by the left-hand *vertical* chain of the figure. This represents the multilevel control mechanism prescribing output targets of current production and rationing inputs. (This aspect of control was discussed on several occasions in Chapters 2–6.) The left-hand vertical arrangement here repeats fig. 5.2 presenting central material rationing, now applied to distribution of investment goods. Of course, only a fragment of the vertical control of current production is to be seen here. Sticking to our example, fig. 9.1 shows the low-level authority that distributes the machines of firms M and N among users. Its superior authorities are not represented here.

The process of carrying out investments and, in this context both the *vertical* multi-level control of current production and *horizontal* relationship of "buyer and seller" were briefly discussed in section 9.4.

What remains to be discussed is the upper part of the right-hand side vertical chain of the figure. What happens about distribution of investment quotas and approval of investment projects between levels 3 and 2, or between levels 2 and 1? This will be the subject of section 9.7. Finally, section 9.8 will deal with the top of the vertical, hierarchical control "pyramid", that is with a few questions of central economic policy.

9.7. The allocator as claimant

Let us therefore begin the analysis of the relationship between levels 3 and 2. The direct superior authority, that is the lower-level allocator for the screw factory – our previously used example – is the Board of the Metalware Industry. The same Board is now seen in another role: not as an allocator but as a claimant.[9] Its fellow-claimants, with whom it competes for investment quotas, are the Board of the Precision Engineering Industry, the Board of the Machine Tool Industry, the Board of the Telecommunications Industry, and so on. The allocator is the medium-level management organ, in this case the Ministry of the Engineering Industry. It is the Ministry's investment quotas that will have to be distributed among the boards.[10]

The sharing of quotas is connected with the approval of the most important of the high-priority investment projects. This is because they may have to be approved at the medium- or upper level, depending on what the law requires. And, even if decision rights were formally held by the lower-level management authority, in the course of negotiations about the division of quotas the question of which specific projects the quotas would be spent on would always be discussed.

Let us recall what was emphasized in section 9.2 when discussing motives for investment hunger: economic management is influenced by

[9]Attention is drawn here to the fact that the order of *discussion* of the book is not necessarily identical with the usual *chronological* order of events. In the discussion we progress from below upwards. In some cases an investment action was initiated from below. In some other cases, however, the decision process starts at the top and passes downwards, although iterations go back to upper levels for possible revision of earlier decisions.

[10]It is only for illustration that exactly four levels are shown in fig. 9.1; calling the medium-level allocator the "Ministry of Engineering Industry" is quite arbitrary, etc. The number of levels of vertical control, the names and functions of institutions in it are different in each country and may also change within any one country over time.

expansion drive at all levels of the hierarchy. The directors of the Board of the Metalware Industry identify themselves with their sector in the same way as the director of the screw factory identifies himself with his factory. One *must* grow. And thus one must fight for investment quotas.

When the allocator changes his role he also changes his attitudes. "Downwards" he was restrictive; "upwards" he is expansive. "Downwards" he refused some of the claims and bargained with those whose claim was approved, trying to keep down cost estimates. "Upwards" he puts in a claim which is larger than he really hopes to get approved, since he expects that part of the claim is going to be "cut down". He does not mind if cost estimates are "overoptimistic" since this may ensure a better chance for approval.

Yet it is impossible to have a perfect "split personality". The lower-level allocator who yesterday fought for larger quotas cannot today unshakeably resist the pressure coming "upwards" to enlarge quotas. Involuntarily a kind of complicity develops between the "final" claimant, namely the firm or nonprofit institution and its direct superior authority against the medium- and upper-level allocators. The lower-level allocator becomes "representative of interests" of units under his management; he "fights" for them. If he passed the downward-biased cost estimate to his own superior authority, he cannot proceed to uncover distortion. If he himself is glad to see as much investment started in his own field as possible, knowing that it will definitely be finished, he will be unable to fight firmly against the thoughtless and dispersed starting of other investment projects.

We do not intend to be repetitious. All that was said about the connection between levels 3 and 2, and about the lower-level allocator's behavior, also holds one level higher up, that is for the connection between levels 2 and 1 and for the medium-level allocator's behavior.[12]

At this point we again wish to consider whether medium- and high-level leaders exert some self-restraint in their investment claims. This might be done not for tactical reasons only, but because they know that investment takes important resources away from consumption and other key targets.

The full dilemma of "investment versus consumption (or other important uses of resources)" appears exclusively at the peak of the decision hierarchy. Top political leaders not responsible personally for any separate partial sphere, such as the chairman of the National Planning Office, and

[11]See Tardos (1972).

[12]When talking at this point about the distribution of quotas at levels (3,2) and (2,1), we focus on the behavior of the institution playing the *claimant's* role (i.e. the one "below"). In Chapter 10 we shall revert to the question of how the institution playing the *allocator's* role (i.e. the one "above") behaves in this situation.

their direct apparatus experience this choice as an *internal* dilemma. They must be fully aware that they are responsible for investment *as well as* for consumption, *and* for the country's foreign trade position, *and* for national defense, and so on. But for everybody else – all those below the top and responsible for some *partial* field – these dilemmas are *external.* As consumers they are, of course, interested in the development of consumption; as citizens they know that the balance of payments and national defense are also important; as experienced economic leaders they can clearly see that investment competes with all those objectives in taking up resources. All that is one way of thinking about it and it is the other way of thinking which becomes active when it comes to dividing up investment quotas. In this alternative way of thinking they feel that they *have* to fight for the development of their own partial field since that is their own direct responsibility. Their social role is much stronger than their individual consumer approach or the impartial understanding of economic interrelations.

9.8. On changes following the economic management reform

After our description of investment-control in the *traditional* management system of the socialist economy, we now say a few words about the situation that followed the *Hungarian reform of* 1968. We do not provide a detailed description and only outline one or two changes.

A remarkable rearrangement of decision competence took place in the *vertical* processes of control. The proportion of investment decided within the firm's own sphere of authority and financed exclusively from its own resources grew – although it still remained rather low. Within the total of financial resources the share of credits to be repaid grew in relation to the so-called "state subsidy". It must be added, however, that in a large sector of investment decisions the role of central authorities (central planners, the central financial and credit system) remained extremely important. Some investment projects are also legally decided by central organs. Or, through granting or refusing state credit or state subsidy they may strongly influence the investment projects which legally are not centrally decided, but which the firm is unable to realize from its own financial resources.

Decision processes for investment no longer follow the scheme depicted in fig. 9.1. In a certain sense control became two-level; the firm applies directly to the central authorities for credit or a state subsidy. Ministries and sectoral organs do not get "investment quotas" to distribute among claimants subordinate to them. Yet they influence the allocation of investments in that they may submit proposals to firms under them, as well as to their superior authorities. As a result of the elimination of the "quota distribution" system ministerial employees are less "double-hearted". They may "fight" more strongly for the development of their own sector since, in accordance with the terminology earlier introduced, they need only play the role of the "claimant" without being at the same time "allocators".

As regards the behavior of the firm, an attempt was made to "internalize" the dilemma of "investment versus consumption". (It has been just pointed out that in the traditional economic management system this is only an external dilemma for the firm's manager.) The firm has a certain freedom of choice in how much of its profit it will spend on its own investment and how much on increasing personal incomes paid by the firm to the employees. The state tries to influence this choice through tax regulations.

The execution of investments is largely influenced by thorough changes in the control of current production. As mentioned in previous chapters, short-term plan directives to firms ceased. Output is not laid down and, with a few exceptions, there is no central material rationing. Thus, *horizontal* relationships between producers–sellers and users of investment goods have come to the fore.

We do not undertake a full appraisal of the situation that has developed as a result of these changes. Experience is not sufficiently clear, and the views of economists differ on numerous questions. We shall only mention one or two ideas that are accepted by quite a few.

It seems that the reform freed the investment sphere from several bureaucratic restrictions. It gave more flexibility to the selection of investments, decisions on the starting of projects, as well as the execution of investment.

The other side of the coin is that *investment tension has remained*, though its intensity, on average over time, may have lessened somewhat. The demand of firms intending to invest for credit and for state subsidy still largely surpasses the amounts available, even now after the reform. Investment plans are taut. The execution of investments permanently hits resource constraints and shortage on the market for investment goods is intense.

This indicates that the reform was not deep enough to eliminate the factors which permanently reproduce tension. The behavior and motivation of firms and of nonprofit institutions, as well as of the lower- and medium-level management authorities, have not changed in regard to investment. Organizational forms have changed, the number of levels in multi-level control has decreased, spheres of authority are divided differently, the order of financing has changed, and so on. Yet *expansion drive as well as an insatiable investment hunger go on asserting their full effect. Investment intentions are still not held back by fear of financial failure, or by feelings of risk. No internal economic compulsion puts a voluntary constraint on investment hunger. The budget constraint for investment is soft – it is even softer than the budget constraint for other kinds of expenditures.* (This will be discussed in more detail in Chapter 13.) *These* are the key areas in which we must seek explanations for the reproduction of investment hunger.

The assertion is empirically testable. Those who doubt it must answer the question whether they have ever seen a manager of a firm who would *voluntarily say no* to an investment possibility offered. Let us imagine the following situation: superior authorities of the firm suggest to the managers that they should carry out an investment. The firm will be granted credit to be repaid with interest; there will be material, machine, and construction capacity. In my opinion the response from the firm to such a suggestion would never be: "No thank you, we do not want an investment credit. We are afraid that the investment might not be a success financially, and we would get into trouble in the repayment of the credit."[13] And, so long as this answer is never heard, expansion drive will always break through. Together with that comes investment tension. And, as explained above, that is sufficient in itself to generate and reproduce a shortage economy.

9.9. The sufficient condition for reproducing investment tension

A few observations about *growth policy* are closely connected to what was said about the economic management reform.

Investment tension is strengthened if central economic policy itself heads expansion efforts. That is what happened in Hungary in the years 1949–52. Central economic policy forced the fastest possible rate of economic growth.[14] Energetic measures were taken "from above" to start the greatest possible number of investment projects with the largest possible dimensions. The aspiration levels of the highest organs in regard to growth were

[13] Private conversations revealed that several Hungarian economists could not even imagine a situation where the supply of investment credits exceeds their demand.

[14] See Berend (1974) and Kornai (1972a, b).

very ambitious. In 1951 the targets of the already taut five-year plan were further increased.

In the distribution of investment quotas heavy industry got an outstanding amount, while light industry, agriculture, infrastructure, and services were driven into the background.

"Growth as fast as possible, at any price" – that is what characterized the expectations raised among the economic managers of firms, nonprofit institutions, and lower- and medium-level authorities. Every new project got official prestige; it was a glory to invest.

In comparison with the years 1949–52 the period around the economic management reform, the 1960s and 1970s, showed important changes. The aspiration levels of the top leadership regarding growth became more moderate; less taut plans were envisaged. The high-level economic management tried to protect resources intended for consumption and other purposes against the suction effect of investment demand.

Some shift has taken place in the intersectoral proportions of investment to the advantage of agriculture, infrastructure, and services.

Official expectations raised among economic managers have also changed. Central authorities do not encourage and stimulate them to invest as much as possible, but would prefer to encourage some self-restraint.

Of all these changes in *growth policy* the same can be said as of the changes *in economic management* described in the preceding section: they could contribute to the mitigation of investment tension in certain years and at certain places. They have not, however, eliminated the underlying causes which bring about investment tension and permanently reproduce it.

Let us sum up the lesson to be learned from the preceding section and the present one. In what follows we shall list the most important factors playing a role in the development of investment tension, that is influencing its intensity. All of them have been mentioned in the literature and in disputes about the subject; therefore, the mere enumeration of them does not contain any new understanding. Opinions diverge, however, in what the specific role of each factor is in the intricate network of causes, effects, and interactions.

We wish to clarify what can be considered a *sufficient cause* of investment tension. We are not concerned with determining which factor has a stronger and which a weaker effect, nor with determining the "chicken and egg" question of which factor appeared first. The truth is that their historical appearance is strongly correlated. We wish to determine whether there is one factor which *alone* would be capable of causing investment tension. Hungarian experience allows us to separate out a "sufficient cause" not only by logical reasoning but also through an analysis of historical experience.

We shall enumerate six factors altogether. They are the most important ones. It is assumed that no other factors have played any important role in creating the investment tension. Two variants will be assigned to each factor. It will be enough to summarize them as headings, since each has already been treated in detail.

(1) *Central growth policy.* (a) Growth at a forced rate; (b) growth at a moderate rate.

(2) *The main proportions in the central allocation of investments.* (a) Agriculture, light industry, infrastructure, services driven into the background; (b) the same sectors are not driven into the background.

(3) *Official expectations regarding investment behavior.* (a) "Glory" of investment activity; (b) self-restraint in demand for investment.

(4) *Decision process of investment.* (a) Strong centralization; (b) partial decentralization.

(5) *Decision process of current production.* (a) Strong centralization; (b) thorough decentralization.

(6) *Internal economic forces restraining investment hunger.* (a) There is no internal power of restraint. Financial risk is not felt by those who ask for and those who carry out investment. Budget constraints are soft. (b) An inner power of restraint exists. Those who request and those who implement investments are aware of the financial risks. Budget constraints are hard.

The above-mentioned variants (a) and (b) summarize abstract cases. No such theoretically pure cases are produced in history. The following train of thought seems to be an acceptable approximation of reality.

First assertion. In Hungary in the years 1949–52 the situation approximated to variant (a) for all six factors.

Second assertion. In the period around the reform the situation with respect to factors (1)–(5) approximated not to variant (a) but to variant (b). It was only for factor (6) that the situation remained close to variant (a).

Third assertion. Investment tensions appeared in both periods.

Conclusion. It is to factor (6) that we must look for the "sufficient cause" of investment tension.

Developments in factors (1)–(5) are of some importance. They may play a strengthening or a weakening role. They may influence the degree of investment tension and thereby generally the intensity of shortage. But the state of factor (6) alone decides whether or not investment tension will be reproduced.

9.10. Investment cycles

In the preceding two sections two *historical periods* of the Hungarian economy were opposed to each other: the years 1949–52 and 1960–70. A *long-term* historical change took place in which the general features of growth policy (growth at a forced rate versus growth at a moderate rate), as well as norms and tolerance limits valid in the control mechanisms of the system have been modified. Over the period as a whole, however, certain features of central economic policy were persistent.

In theoretical analysis *long-term historical change* is sharply separated from the *short-term cyclical fluctuations* taking place within a given period. At such times the system fluctuates around the norms and within the tolerance limits specific to that period. Central growth policy and central investment policy which is part of it may well fluctuate, but only around the trend characteristic of the entire period.

In Hungary *investment cycles*[15] took place in both the periods in question. Cyclical fluctuation accompanied the policy of forced growth within the framework of the traditional economic management system, and the cyclical fluctuation of investment did not stop with the policy of moderate growth within the framework of the postreform economic management system. This suggests that this cycle is caused by underlying factors which have not been changed either by the modification of growth policy, or by the reform of economic management.

In each socialist country in which this cycle phenomenon appears, particular, historically unique reasons *also* play a role in the formation of each cycle. Analysis of these is not one of the objectives of the present book. We are concerned only with factors influencing more or less all cyclical fluctuations. Accordingly, we shall attempt to provide a "stylized", abstract description of a fluctuation.

In the upward-moving period of the cycle investments are on the upswing. Superior authorities approved an increasing number of investment projects; centrally, too, more and more projects are initiated. The execution of the investment projects in progress is accelerated. The upswing lasts until the process hits "tolerance limits" and other constraints

[15]In Hungarian economic research the studies by Bródy (1967, 1969, 1970) stimulated the analysis of the investment cycle. His ideas have been developed further in his new book, Bródy (1980).

In recent years a number of authors have been dealing with the question. Among these works the studies by Bauer and Soós are outstanding. (For references see footnote 1.)

on sudden expansion.[16] Three kinds of tolerance limit are of particular importance.

(1) The trade balance and the balance of payments situations become worrying. No clear rule exists as to where exactly the tolerance limit lies. It is obvious to everybody that the mere fact of a deficit and debt is not worrying, provided it can be assumed that the national economy will in future be able to produce the additional amount needed for debt service. The difficulty arises when economic management *perceives* the volume of deficit and debt as giving cause for worry, and wishes to intervene energetically. This is a well-known phenomenon; it is observable in the government circles of all countries where foreign trade plays an important role. And, if it is felt that a drastic intervention is necessary, the most obvious field will be investment which mainly through imports of machinery, has a severe impact on the balance of payments. If, therefore, economic policy leaders feel that "something is wrong with the foreign exchange situation", their first action will be to restrain investments.

(2) Investment projects "compete" with household consumption for resources. This can be seen directly in a few fields: for example the construction industry may draw labor away from services. In other fields this "competition" is indirect and takes place mainly through the medium of foreign trade. There are a number of agricultural products that households would buy which are also saleable on foreign markets. If exports grow at the expense of the supply to households, additional foreign exchange can be spent on importing machines. Again there is no clear rule as to where the tolerance limit lies. It depends on the actual sociopolitical situation, what level and growth rate of consumption the population is content to accept, and where dissatisfaction begins. And, if there is dissatisfaction, at what point it starts to endanger the stability of the system. It is a historical fact that unrest may be so great that it induces leaders to change economic policy.[18] In such cases the restriction of investments again seems to be the most obvious solution, because this allows foreign exchange saved on machine imports to be used for consumer goods, immediately raising living standards.

[16]The concept of "tolerance limit" was introduced in section 3.2. A further explanation will be given in Chapter 10.

[17]This was well illustrated in the article by Gács–Lackó (1973, 1974) about the behavior of planners. We return to the subject in Chapter 21.

[18]This phenomenon is emphasized in the pioneering study of Olivera (1960) on the cyclical fluctuations of a socialist economy. He writes: "...this generalized social disapproval will subject the planning authority to increasing pressure, urging a change in its allocation criteria...".

(3) As was explained earlier in the chapter, the physical execution of investments constantly hits resource constraints and bottlenecks. With investment tension increasing, these phenomena grow more frequent. There are more and more complaints about shortages, unexpected stoppages, disturbances, repeated corrections of plans, and delays. Those who demanded more investment only yesterday, are embittered today because of the innumerable frictions in execution. Again no clear rule can be given as to where the tolerance limit lies. In any case, such grumbling may reach the point where economic management feels it is intolerable.

All three tolerance limits represent a social *acceptance constraint*. The political leadership, the narrower public opinion of economic managers or the wider public opinion of the population feel that they cannot accept any violation of these limits.

Even if the economy hits only one of the three tolerance limits mentioned above, the "brake may suddenly be pulled". (Sometimes the economy reaches two or even all three tolerance limits at the same time.) Usually it is not just a matter of cautious braking, but what the driver calls: "jamming on the brakes". Quite a few investments are stopped,[19] others radically slowed down. Still others that were to start soon are not even begun. Annual investment targets for the next period are given cautiously. The aggregate volume of investment activity does not always fall but its growth slows down considerably.

This lasts for a time. Slowly, however, the shock effect – induced earlier by the transgression of tolerance limits – dissolves. At the same time opposite signals begin to come in. Firms producing investment goods and services become worried. It is not as if they were faced with such grave selling difficulties as are customary in a demand-constrained economy. But they are troubled by the fact that the volume of unfilled orders has been reduced considerably, and the "queue" has shortened abnormally. One can be less selective in the choice between potential buyers than at other times. The feeling spreads that plans are "too easy to fulfill", that they are "loose". There is too much slack. More could be squeezed out of the investment sphere.[20] This increasingly optimistic spirit suddenly becomes a

[19]It is true but only temporarily. Work will start again during the upswing of the next cycle at the latest.

[20]This phase is illustrated for the late 1950s in the study by Ungvárszki (1976). Investment activity too much reduced must once again be increased. Ungvárszki writes (p. 115): "In the whole of the period it was a repeatedly emphasized aim to restore the earlier, 'normal' rate of investments... The investment policy's aim of restoring that earlier rate, not even defined more exactly, seems to have been handled by leading organs as an axiom that needed no verification; no real justification is to be found, why exactly that 'earlier rate' should be considered 'normal'."

strong determination and a new impetus is given to investment activity. The cycle starts again.

Dynamics of investment policy bear the marks of two different models of control mechanisms: those of control according to *norms* and those of control according to *tolerance limits*. The center changes its policy if the state of the investment sector has moved substantially away from the norm, or if it hits a tolerance limit. Signals are conveyed (in contrast with the cycle in a capitalist economy) not by the cyclical fluctuation of prices, but partly by various statistical "quantity" signals such as data on household consumption, foreign debt and backlogs of unfilled orders, and partly by "voice": warnings and complaints received from economic leaders or from the population.

The suddenness of the interventions and the wide fluctuations between the two types of adjustment ("jam on the brakes" – "step on the gas") make the system fluctuate. This is a phenomenon well known in engineering control theory.

We started our discussion of the investment cycle by drawing attention to the secular changes in the socialist economy and to the shift in the normal state – the average of the cyclical motion – as a result of the change. Now we wish to close our explanation with a related discussion. This time we shall not examine the several versions of the resource-constrained system, but instead we shall oppose the "pure" case of the resource-constrained system to the "pure" case of the demand-constrained system. Although we intend to make an abstract comparison, for the purpose of illustration the capitalist economy of the late nineteenth century and of the first third of the twentieth century could represent the demand-constrained system and the socialist economy functioning in the traditional management system could represent the resource-constrained system.

There are numerous important differences between the cyclical fluctuations of the two systems. The motivations of the participants are different, signals and control mechanisms are different, the role of the state is different and so are social factors. All these differences are outside the sphere of our present analysis; we shall examine only one, namely whether there are characteristic differences in the normal value of shortage and slack indicators.

In the normal state of the classical demand-constrained system there is a considerable amount of mobilizable slack, such as employable but presently unemployed labor, and unutilized fixed capital. This slack is mobilizable precisely because complementary inputs are available. Shortage phenomena exist, but only sporadically. The capitalist cycle takes place

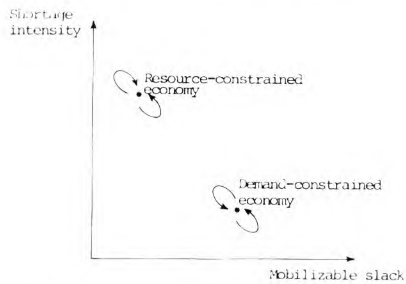


Figure 9.2. Cyclical fluctuation around the normal state.

around this normal state. Even at the peak of the upswing it relatively rarely hits physical resource constraints and at the trough of the cycle slack grows immensely.

The normal state of the traditional resource-constrained economy is very different. In it there is very little mobilizable slack, and intense shortages occur. The cycle takes place *around that normal state*. Mobilizable slack is rare even at the bottom of the down-swing; during up-turn increases in production frequently hit resource constraints. At peaks of the cycle such constraints become even more important. Shortages are considerable even at the lowest point of the cycle, and at its peak they are almost intolerably high in several sectors of the economy.

The difference between the two cycles roughly outlined above is shown in fig. 9.2. Its structure is identical to those given in Chapter 8 and therefore it does not need any detailed explanation. Here, too, vectors composed of shortage and mobilizable slack indicators are each represented by a scalar. Two points, lying far from each other, are shown in the figure: one is the normal state of the resource-constrained economy, and the other one is that of the demand-constrained economy. The cycle revolves around these points.

CHAPTER 10

The investor: Long-term adjustment

10.1. Introduction

The present chapter does not constitute an independent unit but rather is a continuation of the previous chapter. In Chapter 9 primarily the *institutional framework* and the decision-makers' *motivation* were analyzed, while we disregarded what was actually the *content* of decision-making: how investment projects are selected from a set of proposals, and how investment resources are allocated. The answer to these questions is to be given in this chapter.

From the outset, we must make clear that the answer will be far from complete. It would be beyond the scope of this book to give a detailed and comprehensive account of the five-year or fifteen–twenty-year plans of the socialist countries.[1] We shall rather arbitrarily take two or three issues from this extremely far-reaching set of questions, issues closely connected with the main subject of the book.

One "leitmotiv" of the book is the examination of "*quantity*" *adjustment processes* taking place in the economy. After discussing in Chapters 2–8 the instantaneous and short-term mutual adjustment of production and consumption, and of selling and purchase, it is now the turn of *long-term* adjustment.[2] Investment allocation between sectors, regions, and alternative projects is among the most important elements of this adjustment.

[1]See footnote 6 in Chapter 1 for the literature about planning.

[2]At this point we must make a terminological remark. In Hungary the elaboration of the yearly plan is called "short-term" planning, that of the five-year plan "medium-term", and that of fifteen–twenty-year plans "long-term" planning.

In my book I do not use the notion "medium-term". As regards the distinction between "short-" and "medium-term", we follow the usage of microeconomics. The distinction is based not so much on the calendar length of the period, but – at least on the supply side – mainly on the assumption concerning fixed capital. "Short-term" means that we examine processes taking place with *given* fixed capital; "long-term" includes the *change* in the stock of fixed capital.

"Long-term" adjustment is examined in connection with immediately forthcoming investments. Therefore, our time horizon usually does not exceed five years. Problems of the "*very* long term", i.e. analysis covering 15–20 years or an even longer period, are beyond the subject of the present book.

Planning is adjustment – but not *only* adjustment. Its functions are not confined to the coordination of different processes. Yet we shall now focus our attention on that role of planning. And in this respect it is not the methodology of coordination but its *behavioral regularities* that we shall investigate. We consider planning as a special control mechanism in which certain impulses and signals give rise to certain reactions. We want to understand a few of these behavioral regularities.

The other "leitmotiv" of the book – the most important one – is *shortage*. It is natural that from the wide range of problems of investment allocation we should take questions which are – either as cause or effect or as an accompanying phenomenon – closely interdependent with shortage.

One of the most important elements of long-term adaptation is the adjustment of the domestic economy to the expected situation in the external world. We shall not deal with this, similarly to other parts of the book, but shall only refer to it in one or two places.

Most of the chapter is concerned with such phenomena as are present equally in the traditional and the postreform management system of socialist economy. If there is some difference between the two management systems in the questions that are about to be analyzed here, we shall draw attention to it.

10.2. Point of departure: Allocation according to permanent proportions

Let us revert to section 9.7. We discussed there how – in the *traditional* socialist economic management system – the medium-level allocator, for example the ministry, distributes investment quotas among low-level claimants, such as the sectoral management authorities (vertical relationship between levels 3 and 2), or, the upper-level allocator, like the Planning Office, among medium-level claimants (vertical relationship between levels 4 and 3). The factors leading to investment tension in medium- and upper-level allocation were analyzed. The signals and criteria serving as the basis for the quota distribution between claimants at these levels have not been explained.

The point of departure is a very simple rule of thumb: *every claimant should be given the same share of the quota as he received in previous years.* Such a rule would be impossible to apply in practice at the lowest level of allocation because of the indivisibility of investment projects. First the bulk of the investments underway in firms A and B must be finished, and only then is it the turn of firm C. The larger the unit in question – subsector, sector, or maybe a group of sectors under the supervision of a ministry – the more it becomes possible to distribute aggregate investment quotas in permanent shares.

The main attraction of the rule of permanent shares is its simplicity and convenience. A permanently recurrent habitual decision is involved which every organization tries to determine by simple rules of thumb. Allocator and claimant may both refer to a *status quo*. Its danger lies in the same fact: it stiffens proportions and thereby also increases the frictions of long-term adjustment.[3] It was already pointed out in the previous chapter how difficult it is to squeeze unplanned, late-arriving investment claims into investment quotas once distributed. Yet the problem starts before that. It is not easy for the claimant to succeed in obtaining a share greater than the usual one. And, even if he succeeded in convincing the allocator of the justification of increasing his share, the latter would still have to be very energetic in his contacts with other claimants who would resist, referring to their "acquired rights".

In Hungary in the postreform economic management system, medium- and low-level management authorities – as was said in section 9.8 – do not receive investment quotas to distribute downwards along the vertical chain. Yet economy-wide plans contain investment targets in enough detail to influence the proportions of investments and the distribution of investment credits and government subsidies between sectors and subsectors.

How far the tendency of "freezing" the *status quo*, that is stiffening of earlier allocation proportions, prevailed in various periods and various fields can be tested empirically.

Investment allocation is a multistep decision process. If we formulated it as an algorithm, the application of the principle, "to each according to his previous share" would only count as the *first step*. It would be decided in later steps where and to what extent deviations should be made from this principle, and which should be the exceptions that prove the rule. The weightiest factor that may lead to departures from the customary shares is when the allocator perceives that *there are difficulties* somewhere. Let us look more thoroughly into this phenomenon.

10.3. Choice based upon shortage signals

One of the most important signals that influences investment allocation and its deviation from the habitual proportions is information about shortage. In some cases a *preliminary* signal is received: let us begin (or, if

[3]At this point we revoke section 8.6 in which we discussed at an abstract level the rigidity and delays in adjustment as one of the important manifestations of friction.

it is in process already, let us expand) the production of a certain product, otherwise there will be shortage of it. The argument assumes special emphasis if shortage is already present when it is voiced. The more intensive the shortage, the better are the chances of the proposal's acceptance. Investment decision is not infrequently a delayed reaction to the signal: the queue has grown abnormally long, the backlog of unfilled orders piled up, and there are all too frequent complaints on the part of the users.[4]

In certain cases the appearance of shortage is linked to problems treated in section 9.3. The basic investment has been carried out, and shortage signals must be received so that complementary investments can also be executed. Investments have created the capacity for the production of a certain final product, and shortage signals encourage development of the "hinterland" industry, establishing the manufacture of materials and parts necessary for the final product.

Decision is in many cases related to the dilemma of "home production versus import". The product we are short of may be obtained by the introduction or expansion of domestic production, or through imports. Examination of this particular decision problem goes, however, beyond the scope of this book.

Shortage signals influence not only which products should be produced, but also the input combinations, that is the choice of technology. A gradual shift is taking place toward technologies utilizing comparatively less live labor, since managers increasingly feel labor shortage.

On the side of both outputs and inputs "quantity" signals concerned with shortages and future demands are much more effective than economic efficiency and probability calculations using prices. We shall revert to this in Chapter 14.[5]

10.4. Introductory example: Social cost as a function of utilization

Shortage signals, which were discussed in Chapters 2–8 as well as in section 10.3, are based on the consumer's (the buyer's) perception: he cannot accomplish his consumption (buying) intention. Shortage is nothing other than a deviation between an initial intention and its fulfillment.

[4]The article by Gács (1976) illustrates with the example of the construction material industry that the allotment of investments closely follows shortage signals. Given an increasing shortage intensity, the response is a more generous allocation of investments.

[5]We shall discuss in Chapter 13 how far *price* signals influence investment decisions.

Therefore, this form of shortage can only exist where a sufficiently specified intention (consumption or buying intention), against which fulfillment can be measured, has existed in advance.

There exists, however, another, more indirect form of the perception of shortage. For example, many complaints are received at the transport department of a factory from drivers: roads are jammed, driving is increasingly tiring, and the traffic is dangerous. We could say that there is a "shortage" of sufficiently wide and high-quality motor roads as compared to the requirements of the traffic. However, no clear, well-defined, preliminary utilization intention exists ("I wish to use so many roads of such quality"). "Shortage of roads" is expressed by the fact that increasing congestion entails increasingly grave social consequences. This particular form of shortage mainly appears in the service sectors of the economy, or in the service divisions of manufacturing plants.

In this and the following sections we treat this latter form of shortage, and the related behavioral regularities and investment allocation mechanisms. We shall elaborate in more detail the example introduced above, by examining a road connecting two junctions, which is used by a large number of vehicles. At the moment of our examination it is of a mediocre quality, being a two-lane road with a second-rate surface, which is rather worn but still fit for traffic.

Utilization of the road should be measured by several indicators. Utilization is fluctuating in time, varying according to the hour in the day, the day in the week, and the season of the year. And, over a given period, traffic can be measured in several ways, for example according to the number of vehicles passing, or according to their weight, or their length. Let us now take just one representative indicator from the components of the utilization indicator vector; let it be, for example, the number of vehicles passing some point on the road,[6] as a daily average. This is denoted by x and measured on the horizontal axis of fig. 10.1. As an approximation we assume that the larger the number of vehicles passing, the greater the utilization of the road.

Let us consider the *social costs* that depend on the utilization of the road. Some are shown in table 10.1, without aiming for completeness.

If the state of the road is given and its utilization grows, it will be found that the speed of the vehicles becomes uneven: there are frequently repeated stops and restarts, or at least slow-downs and accelerations. This

[6]Let us assume, for simplicity, that every vehicle passing on this road goes from the beginning to end of the road. Therefore, the length of the road covered need not be taken separately into account in the cost function.

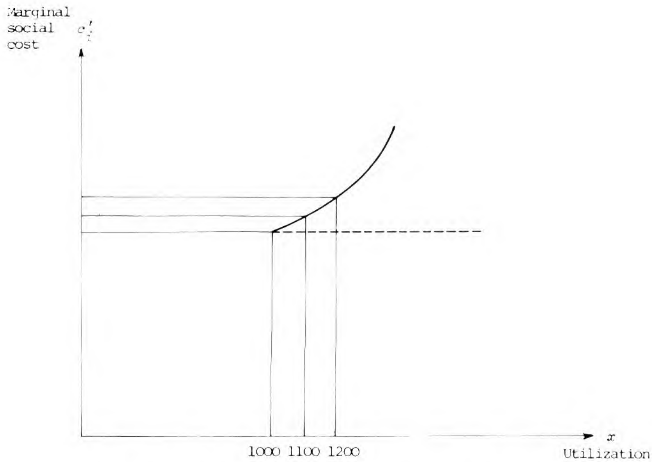


Figure 10.1. Marginal social cost as a function of utilization.

Table 10.1
Social costs depending on road utilization.

Type of costs	Specific features of costs	
1. Fuel consumption of vehicles of firms and nonprofit institutions	Appear in the expenditures of firms and nonprofit institutions in money terms	"Internal" costs from the point of view of the sector of firms and nonprofit institutions
2. Physical wear and tear of vehicles of firms and nonprofit institutions		
3. Time spent on road by professional drivers		
4. Road maintenance		
5. Fuel consumption of private vehicles	Appear in the expenditures of the household in money terms	"External" costs from the point of view of the sector of firms and nonprofit institutions
6. Physical wearing out of private vehicles		
7. Time spent on road by private drivers	Do not appear directly in expenditures in money terms	
8. Nervous tension of drivers		
9. Number of accidents[a]		

[a]A financial "counteritem" appears indirectly in the form of damages paid by insurance companies.

increases fuel consumption, intensifies the wear and tear of the engine, and reduces the average speed of travel. What is more, the road is more intensively used, and its condition deteriorates more quickly. And this further increases fuel consumption and the wear and tear of vehicles. On more congested roads overtaking is more difficult, sudden braking is more risky; nervous tension grows and the number of accidents grows with it. To sum up, all the kinds of cost shown in table 10.1 grow progressively as a function of utilization; in other words, *marginal cost* increases ever more steeply.

The term "social cost" is used, as shown on the table, as a collective term. Inputs considered as costs in a narrower and conventional interpretation, such as fuel consumption are included there. But we also list all other kinds of loss, sacrifice, and burden suffered by society, even if they are not customarily accounted in money terms. Time spent at the wheel by the private driver is a cost for society, as is ill-health resulting from an accident. Such costs are also observable and measurable. For example, the least conventional "cost" is item 8, the nervous tension suffered by the driver. Traffic psychologists can measure this in an indirect way by recording the driver's blood pressure or the functioning of his heart.

We do not aim at measuring artificially in money terms every kind of social cost, according to the methodology of cost–benefit analysis.[7] We do not even raise the question as to how many Forints* the driver's nervous tension in a congested traffic area, or physical fitness lost in an accident are "worth". Every kind of cost is measurable in its specific unit (or units) of measurement. Thus, it might be measured by the total number of accidents, or may be subdivided into different classes according the consequences of accidents. It is noted, therefore, that *social cost is a vectoral concept*.

The vector cannot be drawn in fig. 10.1. Therefore, one component, c_i, is taken for illustration. This is a function of utilization x. In the figure the first derivative of the cost function, $c_i'(x)$, that is marginal social cost, is shown on the vertical axis; as has been mentioned, it is increasing. Let us assume, for example, that if 1,000 vehicles pass the road daily, the total

[7]Thus are called the computations which convert the benefits and costs of some project into imaginary "returns" and "expenditures". Inferences about the profitability of the project might be drawn from the comparison of these "returns" and "expenditures". "Shadow prices" may deviate from actual present or future prices, if that is deemed necessary by the methodology of the analysis, for a better reflection of social benefits and costs. Besides, "shadow prices" may also serve to take into account benefits and costs which have no actual price expressed in money. See Little–Mirrlees (1974), Mishan (1975), and the volume edited by Layard (1976).

*The Forint is the Hungarian unit of currency. (*Editor's note.*)

fuel consumption amounts to 2,000 tons daily; in the case of 1,100 vehicles, fuel consumption is 2,220 tons daily; and with 1,200 vehicles their total fuel consumption will amount to 2,470 tons. Thus, marginal consumption is initially 220, then 250 tons. A broken horizontal straight line is also drawn in the figure. This is a *hypothetical* marginal cost curve which would hold if increased traffic did not entail an increased marginal fuel consumption.

We now proceed to a more *dynamic* analysis. In fig. 10.2 the utilization data on the horizontal axis have time-"labels". The expression $x(t) = 1{,}000$ means that at time t (e.g. on January 1970) traffic was of that volume. And the expression $x(t+1) = 1{,}500$ means that at time $(t+1)$ (e.g. on 1 January 1975) the number of vehicles was 1,500 daily. This shift along the horizontal axis takes place in time *in any case*, independently of the state of the road.

There are four curves on the figure. Which one is valid depends on the maintenance of the road and on investment. Only one curve can be "valid", that associated with the maintenance and investment policies actually carried out; the other curves are hypothetical.

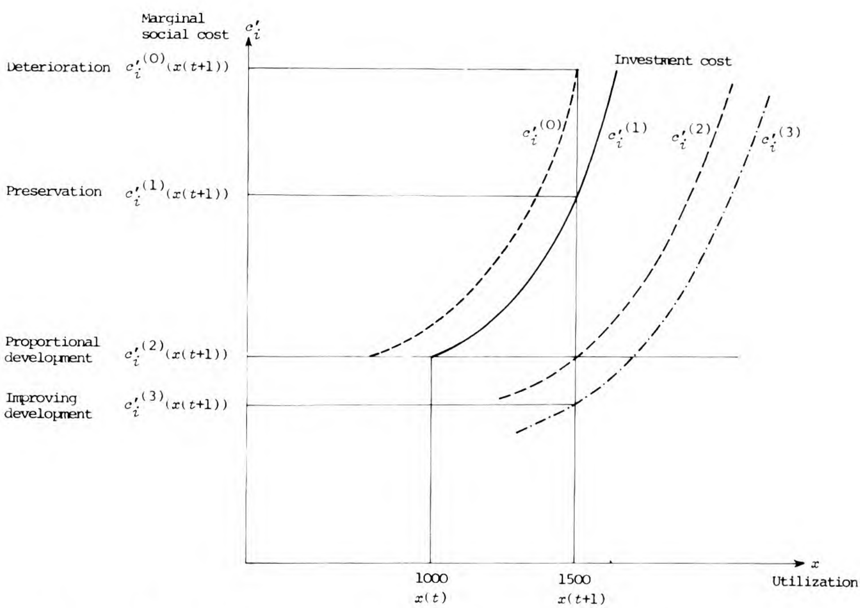


Figure 10.2. Effect of investment alternatives on shifts in the marginal cost function.

We begin our explanation with the curve drawn as a continuous line, which is just a repetition of fig. 10.1. This marginal cost function would remain valid if the state of the road at time t were preserved unchanged by careful maintenance and renovation until time $(t+1)$. This is *investment alternative no.* 1, referred to as *preservation*.

It is worth noticing that, although with alternative no. 1 the quality of the road is preserved at its original standard, marginal cost increases because of the rising utilization of the road. Average fuel consumption has grown, each driver spends longer time on the road, and accidents occur more frequently.

The broken line to the right of the previous curve represents *investment alternative no.* 2. This might mean, for example, that, by the time five years have passed and traffic has grown by 50 percent, a third lane has been added. As a result of this investment, marginal cost associated with the traffic at time $(t+1)$ is – in spite of the increased volume of traffic – exactly the same as with the traffic at time t, for the original condition of the road:

$$\underbrace{c_i'^{(2)}(x(t+1))}_{\boxed{\substack{\text{marginal cost}\\ \text{after develop-}\\ \text{ment at in-}\\ \text{creased volume}\\ \text{of traffic}}}} = \underbrace{c_i'^{(1)}(x(t))}_{\boxed{\substack{\text{marginal cost}\\ \text{before development}\\ \text{at old volume of}\\ \text{traffic}}}}. \tag{10.1}$$

Investment no. 2 has brought about *proportional development*. Development is called proportional if, at a growing rate of utilization, an unchanged marginal cost is secured. On this curve "shifted to the right" the point associated with $x(t+1)$ lies on the hypothetical horizontal line corresponding to unchanged marginal costs.

Of course, at time t the marginal cost for traffic $x(t)$ is an arbitrary starting-point. Efforts may also be made to improve the situation, and this is represented by *investment alternative no.* 3, represented as the curve made up of dots and dashes. This might mean, for example, that the third lane has been constructed with a first-class surface, and that the two original lanes have been resurfaced to the same standard. In spite of increased traffic, marginal cost has fallen below its previous level:

$$c_i'^{(3)}(x(t+1)) < c_i'^{(1)}(x(t)). \tag{10.2}$$

This is called an *improving development*.

And let us now also consider the opposite case, where even the mere preservation of the state at time t is not accomplished, only part of the due maintenance and renovation being done. This is *investment alternative no.* 0 represented by the curve on the left in fig. 10.2. With the increased volume of traffic, $x(t+1)$, marginal cost is higher than if a policy of careful preservation had been pursued:

$$c_i'^{(0)}(x(t+1)) > c_i'^{(1)}(x(t)). \tag{10.3}$$

This investment alternative leads, therefore, to *deterioration*.

Of the four alternatives, nos. 0 and 1 imply *postponement*.[8] With these alternatives increasing traffic leads to increasing marginal social cost. This takes place, as has been seen, not only with alternative no. 0 leading to a definite deterioration in the state of the road, but also with alternative no. 1 preserving the state of the road: the latter one also implies "postponement". On the other hand, alternatives 2 and 3 do *not* imply postponement. They prevent growing utilization from leading to growing marginal costs.

10.5. Temptation for postponement

The allocator who has to decide on road investments is faced with the following dilemma.

On the one hand, alternatives 0, 1, 2, and 3 require, in this order, successively increasing amounts of investment (maintenance, renovation and investment for improvement, in total). It is additional investment that pushes the marginal cost function to the right. On the other hand, to the traffic in the future, $x(t+1)$, the then marginal social cost decreases as we consider the alternatives 0, 1, 2, and 3 in turn.

There will certainly be a number of readers who, having come thus far, will consider this a trivial problem for cost–benefit analysis. Given shadow prices for the various types of social cost, and a social discount rate, they will calculate the optimum alternative in just a few minutes.

Yet there are two difficulties involved. One is that we have no idea about desirable shadow prices. It has just been observed that we do not know the Forint "equivalent" of one hour's nerve-racking drive, one hour's loss of leisure time, or of becoming crippled in an accident.

[8]This discussion is a continuation and completion of the train of thought started by the author's book *Rush versus Harmonic Growth* (Kornai, 1972a, b) on the phenomenon of "postponement".

The other difficulty is that it is not the aim of this book to give advice to the road-building authority as to which alternative it should choose. Our aim is *description*. What is the typical decision made in such dilemmas? And not just description, but also understanding: *why* decision-makers choose what they do. We do not intend to judge whether their decision is right. We want to recognize the causes of the decision.

The following observation can be made.

In numerous fields of the economy there is a strong temptation to postpone due development. Postponement lasts until marginal social costs reach a critical value, the tolerance limit. At that point investment projects so far postponed are carried out, which suddenly lowers marginal social costs. Then again there is postponement, and the process starts once more.

Thus, a partial cycle[9] is formed: postponement overstepping the tolerance limit, investment action and subsequent essential improvement of the situation, postponement, etc. The process is demonstrated in fig. 10.3. We shall revert to the example of the road; the present figure is a continuation of figs. 10.1 and 10.2. On the horizontal axis the increasing utilization of the road is measured at times t_1, t_2, t_3, and t_4. (These are each separated each by periods of several years.) Development of the road had been long postponed until, at time t_2, marginal social cost hit the tolerance limit $\hat{c}_i'$. This was a warning that investment must now begin. This is done in the period $[t_2, t_3]$ and as a result the cost curve is shifted to the right. From time t_3 postponement is going on again.

The heavy curve shows the path of the marginal cost $c_i'(x(t))$ associated with the current actual utilization $x(t)$. The curve fluctuates periodically, reaching the tolerance limit, stepping over it a little, and after the investment project is completed, falling radically, only to rise again in the period of postponement. The lower dotted line shows the minimum marginal cost: the decision-makers do not strive to decrease it below that level.

The figure represents the process in a stylized manner: reality is, of course, never that regular; however, in spite of its simplifications, it helps understanding.

We may ask why the allocator should be prone to postponement. It is not his special mental constitution that makes him open to temptation, but the conditions and the nature of the choice dilemma. The *advantage* of postponement lies in the saving of investment costs, while its *disadvantage* appears in rising marginal social costs. The following factors explain the allocator's behavior.

[9]The attribute *partial* is used to distinguish the phenomenon described here from the *general* economy-wide cyclical fluctuation discussed in section 9.10.

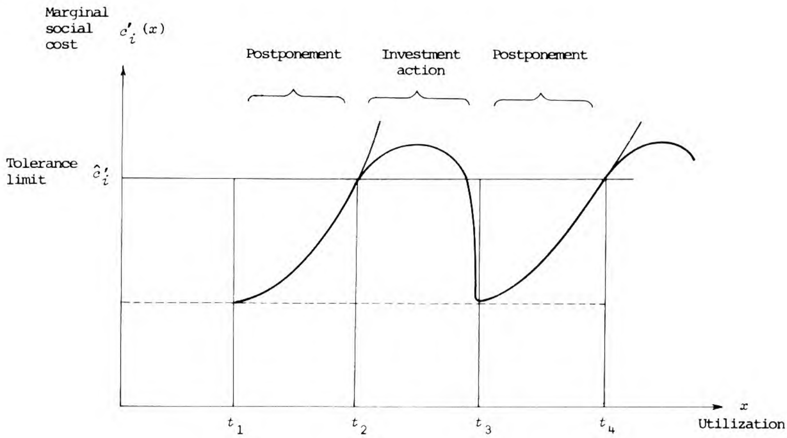


Figure 10.3. Partial cycle caused by postponement.

(1) The advantage is quite *certain*, while the disadvantages of postponement are much more *uncertain*. How much cheaper renovation is than replacement of the road surface and construction of a third lane can be estimated rather exactly. On the other hand, it is difficult to estimate the disadvantages, for who can say exactly, how much petrol would be saved by a third lane, for example.

(2) Advantages are entirely "*internal*", while disadvantages are largely "*external*". The former is *easily measurable*, the latter only *with difficulty*. Savings brought about by postponement are clearly perceivable in the cost estimates of the organizations that would finance the investment. As regards the disadvantages, let us take another look at table 10.1. A considerable part of the additional losses caused by postponement are not expressed in money terms. And measurement is difficult. It has been mentioned, for example, that nervous tension is measurable. It is – but it is not observed by the official statistical service.

(3) Postponement releases resources which *high- or medium-level authorities can dispose of*. On the other hand, the disadvantages caused by postponement appear "*decentralized*". This is a particularly strong factor encouraging postponement. Money, material, and labor released by postponement are "to hand". At the same time, the disadvantages of postponement are dispersed among hundreds of firms, nonprofit institutions, and households, only as small, hardly noticeable items for each one.

(4) The advantage of postponement appears *immediately*, while its disadvantage only *with delay* and *prolonged* in time. This fourth aspect can be described by the habitual formalism of "time-preference", the subjective discount rate of decision-makers being extremely high.

10.6. Tolerance limits

Although the temptation to postpone is strong, the allocator will sooner or later bring himself to do investment since, as has been shown in fig. 10.3, he hits the *tolerance limit*, that is the social acceptance constraints.

The tolerance limit is not a stone wall which makes further progress impossible. Although in fig. 10.3 only a single critical value $\hat{c}_i'$ is given, in reality it is more a *tolerance interval*. There difficulties have multiplied, and therefore those affected by postponement resist more firmly. And although problems as well as resistance exist, even if the allocator ignores them, the tolerance limit will become *effective* only if the decision-maker *perceives* that he has hit it.

Several *signals* warn about the proximity of the tolerance limit. "Grumbling" grows louder. The professional driver complains to his bosses about the bad road, and they will on occasion pass on the complaint to those competent. It may happen by chance that certain people travel on the road who can significantly influence a decision. Press or TV may take up the question. Using the same terminology as before, "voice" serves as signal.

Sometimes it is not just a signal but an *alarm signal* that sounds, indicating that a catastrophe has occurred. In the rush hour, a mass accident might take place on the road, with traffic stationary for hours. Or a section of the carriageway could suddenly become so badly damaged that the whole lane must be closed, causing severe disturbances to the traffic for weeks. Catastrophe then amplifies the "voice" of discontent and complaint.

The tolerance limit appears where the marginal cost curve becomes really *steep*, since that is when problems come thick and fast. Postponement of an investment project entails, as has been seen, increases in many kinds of social cost. To each kind of cost may be amounted some critical value. It is the narrowest constraint that then becomes the effective tolerance limit.

In order to describe and understand any system it is important to know where its tolerance limits lie. This is *one of the basic system characteristics*.

And here we do not mean only tolerance limits constraining postponement, since this is a much more general concept.[10] Tolerance limits depend on many different factors. They depend on the "tolerance" of those afflicted by the negative consequences of the events, the losses they accept without complaint, and how much they are prepared and are able to protest. Limits also depend on the sociopolitical atmosphere, and on whether those for whose ears the "voice" is intended listen attentively to the signals and are prepared to react. In short, the "tolerance limit" is a phenomenon deeply rooted in social relations; it is one aspect of them.

10.7. The method of "putting out fires"

It would be convenient to settle the problem of postponement by the simple requirement that the planner should be intelligent and foreseeing. He should not wait until the marginal social cost reaches its tolerance limit, but should execute the required investment project before that.

That is all right, but the whole affair does not depend solely on foresight. In our example we have talked so far only about a single, randomly selected project (construction of a road section). Let us assume that this road section is of nation-wide importance, and so that investment is within the competence of the main road superintendent authority. This authority has a substantial backlog of postponed road construction tasks, and the same can be said about its superior authority disposing of all transport investments: there are also postponements in the development of railways, air transport, the bus network, and shipping. And transport is, of course, an illustration selected arbitrarily; similar things could be said about numerous other sectors.

And now let us recall the subject of sections 9.5 and 9.7, namely the tension present in the medium- and high-level allocation of investments, which is *inevitable* under the circumstances there described. Demands are high and investment resources are scarce. And the greater part of resources is already committed for the continuation and completion of investments already started. There are hardly any *free* investment resources which one can really dispose of. And if in this situation some claimant has a problem, these few resources must be granted to him. It is necessary "to put out the fire": *investments must be started first of all in the field where, as a*

[10]In Chapters 3, 5, and 9 of the present book tolerance limits already appeared and, of course, these examples are far from exhausting the forms of occurrence of this phenomenon.

consequence of postponement, marginal social cost has hit the tolerance limit.

The mechanism of "postponement, putting out the fire, postponement" forms an economic state similar to that of the man in debt who owes money to many people but tries to maintain his credit worthiness. Therefore he always pays the debt that has expired and cannot be prolonged any further. And, if there is no other way, he borrows money again in order to settle his due debts. Thus, indebtedness is permanent and there are always debts expiring and not postponable.

The analogy leads us to the *foreign trade* aspect of the problem. In the context of the "postponement, putting out the fire" cycle it is important to distinguish between importable and nonimportable goods.[11] The former are mostly products (discussed in section 10.3), the latter mostly services. In railway communication, for example, track and rolling stock are complementary inputs, and both are indispensable. The permanent way must be built at home, so that in this connection all the problems arise that were mentioned in the road example. If, on the other hand, domestic wagon manufacturing cannot satisfy home demand, further domestic development can be postponed without the railway's marginal social costs rising, provided that wagons are imported in sufficient quantity. It is true that this may increase foreign indebtedness, presuming the trade balance and the balance of payments was already negative. And what was said symbolically, by analogy, about a postponed but eventually expiring "stock of debts" (investment projects), now holds literally for the total of foreign debts. *Domestic "debt" can be transformed into foreign debt and vice versa*, not always directly, of course, but at least indirectly. Let us assume that both in sector A providing some nonimportable service and in sector B producing importable goods postponement occurred in recent years, and in both there is an imminent danger of hitting the tolerance limit. Yet free investment resources suffice only to begin a project in one of the two sectors. The following two choices are both feasible. Action might begin in sector A, while imports complementing the domestic production of sector B must be increased. The result is that domestic "debt" has not increased, but foreign debt has. Or, investment begins in sector B, while it is further postponed in sector A. The result is that foreign debt does not grow, but the domestic "debt" of postponements, increasingly hard to accept, does.

[11]It is the usual "tradable–nontradable" distinction that we apply here. This is to show whether the product in question *may* be, in normal conditions, an object of foreign trade, and not whether, in the existing production and foreign trade structure of the country, it is in fact imported or exported.

While tolerance limits are associated with each form of domestic postponement, the amount of foreign debt also has its tolerance limits, as was explained in the preceding chapter.

Now after this digression on foreign trade let us revert to our original subject, namely the mechanism, "postponement, putting out the fire". Putting out the fire creates a vicious circle. It pre-empts the available uncommitted investment resources for the fields (subsectors, regions, etc.) in which problems are accumulating. Therefore nothing or hardly anything is granted to areas in which the situation is still tolerable. There the utilization of fixed capital is constantly on the increase; among other things, quantity drive and the continuous expansion prevailing everywhere in the economy promote this increase. Postponement begins, which sooner or later leads to difficulties there as well. Exactly this – the overstepping of their own tolerance limits – will give them the right to obtain investment resources not yet committed. *Thus, "putting out the fire" permanently reproduces itself.* We talked earlier about the interaction and vicious circle of shortage and investment tension. One of its sub"circles" or partial phenomena is the cycle, "postponement, putting out the fire". As long as shortage and investment tension exist, postponement and the "putting out the fire" method *must* also exist. Expert and farseeing planning may help to some extent, but it seems that it cannot fully eliminate the problem. This is, by the way, an assertion to be checked: it must be tested by experience how far the hypothesis regarding continual self-reproduction of postponement and its concomitant of putting out the fire stands up in practice.

The process described in sections 10.4–10.7 can be considered a *special case of "quantity" adjustment*. It exists not only in a socialist economy, but in every social system. It is a particular form of long-term adaptation by investment projects, adjustment by the method of "putting out the fire". *The signal system of the feedback mechanism involves grumbling, complaints, or perhaps catastrophe. The reaction to the signal requires concentrating uncommitted investment resources in the area that has hit its tolerance limit.* This may be a painful form of progress, but not "irrational". Its regularities are quite clear and can be explained, while experience shows that it is *viable*.[12] Although with delay, it fulfills a considerable part of the role of long-term adjustment.

[12]The method of "putting out the fire", as a control mechanism, can be formalized. Using a few strong simplifying assumptions, Radner and Rothschild elaborated models for its description. (See Radner, 1975; Radner–Rothschild, 1975; and Rothschild, 1975.) These models prove that such a control mechanism is viable, permitting the survival of the system.

Control mechanisms reacting to certain "critical values" arise in many different processes. This is demonstrated for the case of inventories in the study by Farkas (1976).

After all this, we can revert briefly to the assertions of section 10.2. It was pointed out there that the starting-point for investment allocation is the conservation of previous shares. One of the most forceful ways of breaking out of this comfortable rigidity is by "putting out the fire". A larger than usual amount of resources must be spent where problems have accumulated intolerably – even if sectors and regions in which the solution to problems can be delayed receive less than usual.

CHAPTER 11

Employment

11.1. Introduction

One of the basic historically important achievements of the socialist economy is full employment. Not only does it reach a high level of employment but, once having reached it, firmly guarantees it. The present chapter analyzes the processes that led up to full employment and which guarantee its maintenance as well as various phenomena associated with these processes. Capitalist and socialist economies will be compared, but the subject of the chapter does not necessitate – at the rather abstract level at which we discuss employment – the separation of the traditional and postreform systems of economic management in the latter.

Now let us introduce the preliminaries of the theory. It will be clear to the reader how much my train of thought owes to Marx.[1] His recognition of the role of "relative overpopulation", connecting the question of employment to the *secular* analysis of accumulation, is one of the most important contributions Marx made to economic theory.

Keynes and the macroeconomic thinking that has evolved from his approach emphasizes the *short-term* control of an advanced capitalist economy.[2] This led to important practical economic policy conclusions, while it pushed the secular historical view into the background. The latter came to the fore mainly in a roundabout way through the study of *developing countries*. In many Asian, African, and Latin-American countries similar processes are taking place today as in the England of 100 to 150 years ago, witnessed by the classical economists.

[1]See first of all *The Capital* by Marx (1867–1894b). From later Marxist literature we mention works by Luxemburg (1913) and P. Erdős (1976).

[2]See, first of all, the *General Theory* by Keynes (1936a, b), and studies by Hicks (1937, 1974a, b) and Phillips (1958). For a summary see Branson (1972).

Marxist criticism of Keynes's theory is given in the Hungarian literature by P. Erdős (1966, 1971) and Mátyás (1973). The monetarist–neoliberal criticism is to be found first of all in Friedman's works. (See, for example, Friedman, 1975 and 1977b.)

To understand the employment problems of socialist countries some essential facts are supplied by the literature on labor planning.[3]

In this chapter we carry out macroeconomic analysis, mainly in terms of aggregates. We shall primarily examine long-term processes, though we shall touch upon the problems of short-term and instantaneous adjustment as well.

11.2. Classification

Several groups of people will be distinguished, including the population able to work, those employed, and so on. Even if between two dates the number of people in one of the groups does not change, the composition of persons may change as a consequence of demographic processes; people enter and leave each group continuously. We now disregard the effects of demographic processes on the composition of the groups. In general, demographic effects will be treated only at points where it is indispensable to the point of view of our argument.

We seek a *logically* unambiguous classification in the present theoretical framework, but it is not our task to submit proposals for practical implementation. In the empirical testing of our hypotheses the exact borderlines of groups will have to be defined, of course, but that is to be done by experts on labor statistics.

Population able to work. The number of people in the group is $L(t)$. This includes those who, considering their age and condition of health, can work for a firm or nonprofit institution.

Boundaries of the category can be described by objective parameters (age limits, health conditions affecting working ability); but *where* these limits should be drawn cannot be decided by "biological" criteria. Actual limits are based on legal and moral norms accepted by society.[4]

Since it is stochastic magnitude determined by social norms, we shall use the value of the *trend* fitted to the time series $L(t)$ at time t (denoted by $\tilde{L}(t)$) later on.

[3]See, for example, J. Timár (1964) and Kovács (1974). See also Ellman (1979) on full employment in socialist economies.

[4]In a long-term analysis it would cause a problem that the limits of "working ability" socially accepted and considered normal might not be the same at the end of the period under examination as at the beginning. The indicators to be discussed can be interpreted only if the *definition* of the group's limits is the same for the whole period of analysis. In such cases, therefore, the social norms valid at the beginning or those valid at the end of some average overtime must be used as a basis as a simplifying assumption; but in any case there must be a definition holding throughout.

Labor supply. The number of people in the group is $S(t)$. This is the group of the population willing to take on employment.

Employed population. The number of people in the group is $N(t)$, and it includes all who do wage-work in the sector of firms[5] or nonprofit institutions. Therefore, the fisherman and hunter working with their own tools, the smallholder working his land together with his family, the craftsman or small shopkeeper working with his family in his own shop do not belong to this group; nor do housewives. This category does not include every person working, but just those working for wages in formal organizations, that is in firms or nonprofit institutions.

Unemployed population. The number of people in the group is $U(t)$. Its definition is

$$\underbrace{U(t)}_{\boxed{\text{the unemployed}}} = \underbrace{S(t)}_{\boxed{\text{labor supply}}} - \underbrace{N(t).}_{\boxed{\text{the employed}}} \tag{11.1}$$

Population used to employment. The number of people in the group is $H(t)$ and it includes those who have already been employed in the sector of firms and nonprofit institutions, and are therefore used to the circumstances concomitant with it. This group will play a central role in our analysis. Employment is not just a market transaction, work for a wage. Someone who has been a housewife and is now a worker or clerk has changed her *way of life*; as is someone who previously cultivated his own land and is now doing wage-work on a large-scale farm or an industrial plant. In the overwhelming majority of these cases *this far-reaching change in the way of life is permanent and irreversible.*[6]

From individual interviews sociologists would be able to judge clearly about the majority of the questioned, whether they belong to this group,

[5]By the word "firm" we mean not only a medium- or large-scale plant, but also a small-scale one using hired labor.

[6]In standard microeconomics variables can move – within their constraints – freely in both directions. Either demand or supply may decrease after it has increased earlier, or vice versa. This is usually the case in the short run, in the case of marginal changes. But the same does not hold *in the long run*, with the *large* increases or decreases of variables characterizing more fundamental processes. *Asymmetrical situations* mentioned repeatedly in this book are *partly the results of such historically irreversible processes*. Sticking to our example: on the demand-constrained labor market asymmetry (the "buyer": the employer is dominating) develops, among other things, because – as we shall see – the poor man in the village who has become a wage-worker in the town will not go back to the village even if the position on the labor market worsens.

and some difficulty would only arise in border cases. A picture could be formed about the proportion of those who have worked at least one or two years in the sector of firms and nonprofit institutions and still carry on (or would like to carry on) this way of life. The inevitable uncertainties of measurement also justify our neglect in the analysis of the instantaneous value observed at time t. Instead, the value of the *trend* fitted to the time series, denoted by $\tilde{H}(t)$, is used.

It is assumed that those used to employment are always willing to take employment, that is they are part of the labor supply. However, others might also be trying to obtain employment. These form the fifth category.

Population not used to employment, but willing to take a job. The number of people in this group is $J(t)$, and it includes those leaving the old way of life only temporarily. For a time they work in a firm or in a nonprofit institution but, for some reason – family circumstances, or disappointment of their expectations – they leave it again. They may stop even earlier than that, if they searched for employment for a while without success, and then gave up. This group is defined by the following relationship:

$$\underbrace{J(t)}_{\boxed{\substack{\text{population}\\ \text{not used to}\\ \text{employment,}\\ \text{but ready}\\ \text{to take a job}}}} = \underbrace{S(t)}_{\boxed{\substack{\text{labor}\\ \text{supply}}}} - \underbrace{H(t)}_{\boxed{\substack{\text{population}\\ \text{used to}\\ \text{employment}}}}. \qquad (11.2)$$

The preceding classification helps us to construct two indicators. One is h^*, the *normal participation rate*:

$$h^*(t) = \tilde{H}(t)/\tilde{L}(t). \qquad (11.3)$$

This is, therefore, the ratio between those used to employment and the total number of population able to work. It does not indicate the instantaneous state of the labor market, since both numerator and denominator are trend values. It relates the numbers of people in two groups which are bounded by social conventions, customs and norms, summarizes the "change in the way of life", that is the historical process that ties a considerable part of the population to the sector of firms and nonprofit institutions, almost irreversibly.

The second indicator is the *employment rate n* in the customary sense, or its complement, the *unemployment rate m*:

$$n(t)=\frac{N(t)}{S(t)}; \qquad m(t)=1-n(t)=\frac{U(t)}{S(t)}. \tag{11.4}$$

employment rate — unemployment rate

In contrast to the indicator h^* expressing the actual state of a long-term process, indicators n and m represent an instantaneous state in the short-term movement of the labor market. Nearly every macroeconomic model dealing with employment in advanced capitalist countries concentrates on the indicators n and m and does not deal with h^*.

The definitions make it clear that the normal participation rate h^* need not coincide with the intertemporal average of the employment rate, $n(t)$. The former is a *sociological* category (what percentage of the population of working age is accustomed to employment), and the latter a *labor market* category (what percentage of the labor supply succeed in finding employment).

11.3. Demand-constrained labor market

In this and the following sections we contrast two "pure" cases of labor market and employment. One is the *demand-constrained*, the other the *resource-constrained* labor market. We shall talk later about the historical transition that leads from one to the other.

The "classical" case of a demand-constrained labor market is a backward country in the historical period when the capitalist sector is pushing forward, but has not yet reached a high degree of development as it is understood today. Not a "pure" case, but not very far from it is the advanced capitalist economy, assuming that its growth is rather slow in the long run, no powerful Keynesian governmental intervention exists in it, and the influence of trade unions is not strong. That was more or less the situation in most capitalist countries between the two world wars.[7]

[7]The labor market of the capitalist economy in the period following the Second World War cannot be considered a pure case of a demand-constrained market.

Although we can refer to a variety of historical realizations, the "pure" case is a *model* which abstracts in many respects from the various actual circumstances of the different countries.

(1) The most important characteristic property of the demand-constrained labor market is that the normal participation rate h^*, although growing in the long run, always remains low. The adjective "low" is, of course, relative; it will acquire meaning exclusively in comparison to the resource-constrained case. There, as we shall see, the rate h^* is much higher. In an empirical description this will stand out particularly clearly for those countries whose transition from a pure demand-constrained labor market to a pure resource-constrained labor market was comparatively rapid. (For example, Hungary in the 1930s versus Hungary in the 1970s.)

As has been mentioned, h^* may also grow over time in the pure demand-constrained labor market but, despite the growth, there is always *potential reserve labor*.[8] This group will be analyzed in detail in the next section; here we only anticipate its interpretation. The population not used to employment but able to work constitutes the "full" potential reserve labor. Part of this, however, is not mobilizable. In other words, the value of the rate $h^*(t)$ can never reach 1, but it has an *upper tolerance limit* denoted by $\hat{h}$. The definition of potential reserve labor, $Q(t)$, is then as follows:

$$\underbrace{Q(t)}_{\boxed{\text{potential reserve labor}}} = (\underbrace{\hat{h}}_{\boxed{\text{tolerance limit of the participation rate}}} - \underbrace{h^*(t)}_{\boxed{\text{actual normal participation rate}}})\ \underbrace{\tilde{L}(t)}_{\boxed{\text{population able to work}}}. \tag{11.5}$$

The main characteristic of an economy whose employment is demand-constrained is that it always has *large* potential reserve labor. Only part of this appears in the usual unemployment statistics. The number $J(t)$ as defined in (11.2) is always much smaller than $Q(t)$. The rest of potential reserve labor (i.e. $Q(t)-J(t)$) is not "unemployed", just potentially employed, such persons as might become employed in different social circumstances. Their presence, however, fundamentally affects the labor market. *It is primarily the huge weight of potential reserve labor* (and only in the

[8]"The industrial reserve army, during the periods of stagnation and average prosperity, weighs down the active labour-army, during the periods of over-production and paroxysm, it holds its pretensions in check. Relative overpopulation is therefore the pivot upon which the law of supply and demand of labour works." Marx, *The Capital* (1867–1894b, vol. I, p. 639.)

second place unemployment in the strict sense of the word) *that makes the labor market a "buyers' market"*.

(2) Although it is only a secondary factor in shaping market conditions, unemployment of course plays a very important role in explaining the situation. Let us break down the phenomenon of unemployment into components, based on its specific causes and its behavior over the business cycle.

In the literature on employment the expression "frictional unemployment" is well known. The word "friction" is used in the same sense as in Chapter 8, where the indicator vector w was introduced to measure frictional phenomena. Employment of labor is, among other things, an allocative process, for which everything said in Chapter 8 about the frictions affecting such processes still holds. Even if the buyer (firm or nonprofit institution searching for labor) had a buying intention, and the seller (the person intending to take on a job) had a selling intention, it is uncertain whether they will meet. It depends on whether instantaneous demand and supply correspond to each other in every respect and whether the firm or nonprofit institution wants the particular person – considering his professional training or even political attitude – who seeks the job. And, conversely, are the job, the working conditions, the bosses, and location satisfactory to the person considering the offer of employment? Can they agree on wages? And, if the answer is yes to all these questions, do they know of each other, are they mutually informed about each other's intentions and conditions? Answers to these questions are never exclusively positive or negative, but show a definite statistical distribution. *There always exists simultaneously partial labor shortage and partial unemployment in every system.*

In fig. 11.1 we revert to the scheme of representation[9] introduced in Chapter 8. The horizontal axis represents *labor slack*. This is a vectoral category, and we shall discuss later its various components. Here only one component is indicated, the unemployment rate $m(t)$ in its usual interpretation, defined in (11.4). The vertical axis represents *labor shortage*. This is also a vectoral category, and its other components will also be treated later. Only one indicator is shown here, z_L, the ratio of vacancies to the total number of jobs. This means the proportion of jobs for which firms and nonprofit institutions would be prepared to employ someone if they

[9]We can repeat at this point what we already stressed in Chapter 8: in fig. 11.1, as well as in other similar figures, the shape of the iso-friction curves is arbitrary. It is only the signs and a few constraints on the trade-off between shortage, slack and friction that have been clarified by our analysis, but we do not know the exact shape of the friction functions.

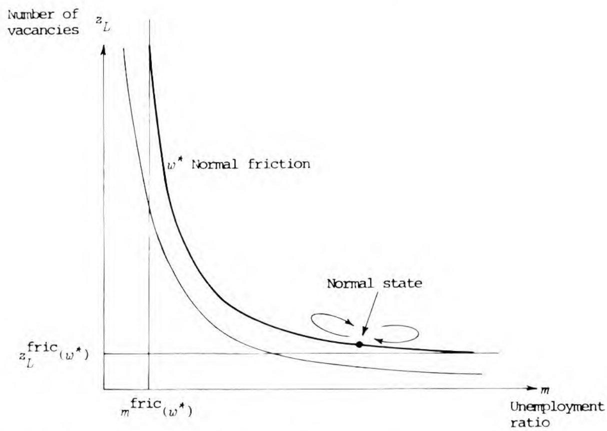


Figure 11.1. Normal state of the demand-constrained labor market.

could find a suitable person. Finally, as the third variable an appropriate friction indicator from the vector $\boldsymbol{w}$ is applied. This may be, for example, the extent of "search", either the number of places visited by the unemployed before finding a job, or the search time, and so on.

In the system under examination – with given normal friction – the iso-friction curve drawn as a heavy line is valid; it approaches asymptotically a vertical straight line on the left and a horizontal straight line on the right. The vertical straight line cuts the slack axis at $m^{\text{fric}}(\boldsymbol{w}^*)$. This is the minimum rate below which unemployment cannot go, for as long as the labor market operates with the normal friction $\boldsymbol{w}^*$. The horizontal straight line cuts the shortage axis at $z_{\text{L}}^{\text{fric}}(\boldsymbol{w}^*)$. This is the minimum below which shortage cannot fall, as long as the normal friction $\boldsymbol{w}^*$ prevails.[10]

The iso-friction curve may shift either upwards, or – as shown in fig. 11.1 by the thin line – downwards. For example, it may be shifted downwards by a better organization of information in the labor market, improvement of the material conditions of labor mobility in housing and transport, and so on. This will be treated in more detail in connection with the resource-constrained state. It is certain, however, that the frictionless state $\boldsymbol{w}=\boldsymbol{0}$ is merely an abstract point of reference in our model and cannot exist in reality. Therefore, "full employment" with $m=0$ can never occur, nor can $z_{\text{L}}=0$, the perfect elimination of labor shortage. We use the expression "full employment" more in a figurative sense, as will be explained later.

[10]Our magnitude $m^{\text{fric}}(w)$ is not identical with the natural rate of unemployment as discussed by Friedman and Phelps, but smaller. See Friedman (1975, 1977b), also Phelps (1970a, 1970b).

In fig. 11.2 we demonstrate, by means of a fictitious time series, the development of unemployment in the "pure" demand-constrained case. The horizontal straight line represents the minimum unemployment rate below which the actual rate cannot go, since frictional phenomena would prevent it from doing so. The actual process of the business cycle determines how near actual unemployment can be to the minimum resulting from friction.

Let us denote by m^* the intertemporal average of the time series $m(t)$: this is the *normal unemployment rate* of the system. The second characteristic feature of the demand-constrained labor market is (besides the low normal participation rate h^* and large potential labor reserve) that the normal unemployment rate considerably exceeds the minimum rate caused by friction; not even at the peaks of the cycle does it fall to the minimum level. That is when we talk about *chronic unemployment*. The intensity of this is measurable by the difference $[m^* - m^{\text{fric}}(w^*)]$, as well as by the difference between the actual minima of the time series and the minimum level explained by friction.

If we now look back at fig. 11.1 we can see the point representing the normal state of the labor market. (In this figure this point is placed in three-dimensional space; in reality, however, it is in a space spanned by three vectors representing labor slack, labor shortage, and friction on the labor market.) The characteristic locus on the figure reveals considerable friction, low but on average positive labor shortage, and (again on average) high unemployment. *Cyclical fluctuation takes place around the normal state*. In the figure this is shown by loops moving away from the normal point but returning to it again.

(3) The third characteristic property of the demand-constrained labor market is that even *some of those used to employment are unemployed*. It is

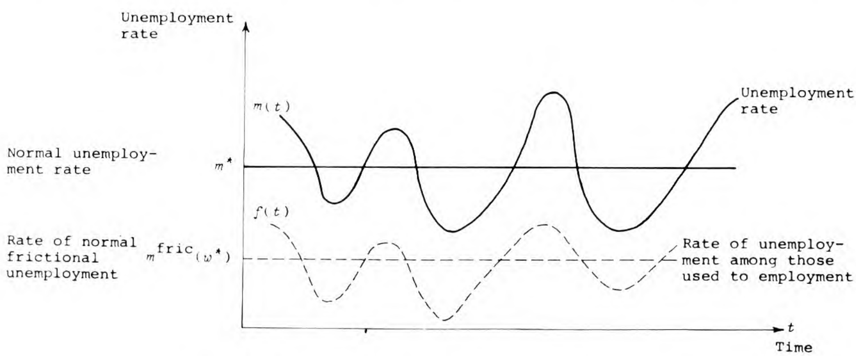


Figure 11.2. Movement in time of the unemployment rate.

true that in most cases they have a higher chance of finding employment than beginners *or* occasional experimenters, but they are not exempted from unemployment, either. Let us denote by $f(t)$ the ratio of those used to employment but instantaneously unemployed to the total labor supply. In fig. 11.2 the time series $f(t)$ is also drawn for illustrative purposes. At the peak of the cycle this particular unemployment ratio might be close to zero, while in the troughs it grows considerably. Every member of this group lives with the possibility of unemployment. The sense of danger is increased further by the existence of the large potential reserve labor. The firm or nonprofit institution can – perhaps at some disadvantage or by forced substitution – always replace a person. Although the firm or nonprofit institution may have need of his expertise and practice, the reverse is even more true. Exactly because a person may be used to employment, he feels that he has little other alternative, and this increases his defencelessness.

The *labor market is demand-constrained, because production is demand-constrained*; its expansion is restricted by the selling opportunities generated by effective aggregate demand. (Let us remember that we are discussing the pure case of pre-Keynesian capitalism, with no artificial expansion of aggregate demand.) It would go beyond the scope of the present book to examine *why* production is demand-constrained. Here we only emphasize the fact because we shall refer to it in later comparative analysis of the socialist economy.

We have stressed three main properties for the pure case of a demand-constrained labor market: large potential reserve labor; considerable chronic unemployment which is not absorbed even at the peaks of the cycle; and even those used to employment are threatened by unemployment. All this accounts for the existence of an asymmetrical state, with *the demand-constrained labor market being a buyer's market*. It is one of the basic ideas in the book that markets cannot always be described as neutral meeting places of two equal forces, namely mutually symmetrical demand and supply curves. It often happens that, to use an expression from sport, one of the parties plays on its "home ground". Under certain conditions, power relations may be balanced. It is, however, quite often the case that the normal state of the market or its intertemporal average situation is characterized from the outset by the superiority of one or the other party, so that one or other kind of asymmetry is prevailing on the market.[11] The most important features of this situation will be lost if we represent the

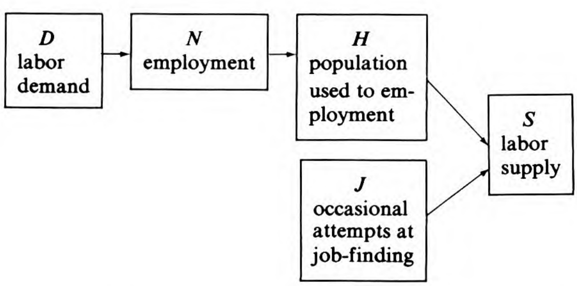
[11]This asymmetry is trivially self-evident for the Marxist economist when he thinks of the capitalist labor market. A careful examination of this analogy can help the economist brought up in Marxist thinking to understand how asymmetry also appears in other markets and in other systems.

labor market by a "Marshallian cross", a downward sloping labor demand curve, and an upward sloping supply curve, both considered as functions of wages. Although there is some truth in such a representation, since it exhibits one of the short-term relationships, the view of the two mutually symmetrical curves obscures the decisive feature of the situation, namely its asymmetry. It is the employer, i.e. the "buyer", who dictates; the employee, i.e. the "seller", is in a defenceless position.

Let us look back at fig. 11.1. Shortage is low, about at the minimum caused by friction. Slack, on the other hand, is large, and the point of the normal state is shifted towards the lower right-hand corner of the figure. This symbolizes, in this mode of representation, the shift in power relations to the buyer's advantage and the seller's disadvantage.

Under such conditions most of the burden accompanying the allocation process is charged to the seller. It is people wanting a job that engage in *search*, and queue up if there are many applicants; they will have to *wait* when there is no immediate offer of a job; they may have to accept *forced substitution* in taking a job below their talent and training.[12]

Important though the role of wages may be in the short-term movements of labor market, they do not belong to the subject of this chapter. They will be treated in Part II, Chapter 16. At this point, we only state the well-known fact that the average wage calculated at the macroeconomic level (both real and nominal wage) is rather rigid, and the supply as well as the demand for labor are rather inelastic with respect to wages. That is one reason why "quantity" adjustment dominates.[13] Although there is interaction between supply and demand, the causality has a definite direction, represented in the following scheme:



(11.6)

[12]We have gone over the same forms of action as were treated in Chapter 4. There they were features of *purchases* and not selling, which weighed on the buyer of inputs for production, in the seller's market of the resource-constrained system.

[13]There one of the central ideas of Keynesian employment theory is touched upon. On this subject, see Leijonhufvud (1968) and Hicks (1974a, b) who discusses the crisis in Keynesian theory.

In the demand-constrained labor market demand is the point of departure. It is the effective constraint on employment (in contrast, as we shall see, to the resource-constrained system in which labor supply will be the effective constraint). It is demand that basically decides employment, so D determines (more or less) N. Actual employment over a long period "accustoms" the worker of the firm or of the nonprofit institution to do wage-work. And, once he is accustomed to it he will (with a slight number of exceptions) be among those wishing to sell their labor-power. What is more, they are joined by those of the potential reserve labor who hope for employment. Their appearance can also be considered as a function of the expected labor demand.

The relative rigidity of labor supply, its wage-inelasticity and, together with this, the decisive role of "quantity" adjustment are partly explained by the fact that H is determined to get a job. *Past* labor demand (having determined past employment) ultimately determines the present situation by "accustoming" the *present* labor supply to the market conditions.

The scheme (11.6) describes the *long-term* adjustment of labor supply to demand. In the short term, and even more in instantaneous adjustment a complex interaction between the variables in the scheme is involved. At this point we merely wish to draw attention to the long-term processes.

11.4. Resource-constrained labor market

The situation of a modern capitalist labor market differs in many respects from the "pure" case described in the preceding section. However, its analysis is not the concern of this book. We shall now make a major leap in historical time and turn our attention to the "pure" resource-constrained case. Later we shall go back in time, because we also intend to examine the period preceding the "pure" resource-constrained state. Yet at this point it will facilitate our reasoning if we contrast the two polar cases.

(1) *The main property of a "pure" resource-constrained labor market is that the normal participation rate is high and has reached the tolerance limit:* $h^* \sim \hat{h}$. *The sector of firms and nonprofit institutions has fully absorbed the potential reserve labor* Q. Let us consider its main components, which will enable us at the same time to understand the tolerance limit $\hat{h}$.

(a) The most important component of potential reserve labor consists of *the independent small-scale producers and the self-employed.* In Eastern European socialist countries the most populous stratum of this group was that of smallholders and middle peasants. The bulk of the labor represented by this stratum has been absorbed in this region (with the exception of Poland and Yugoslavia, in all Eastern European socialist countries) by the sector of firms and nonprofit institutions, the former including cooperatives. This book cannot give a historical account of how the transformation of economic ownership took place in the Eastern European socialist countries. Here it is enough to point out that its concomitant was the flow of labor into the cooperative and state farms of agriculture, as well as into firms and nonprofit institutions outside agriculture. Although the basic trend was determined by this transformation in ownership, the process is not complete. There exist the so-called household plots in agriculture, and their labor requirements present a countertendency to the absorption effect of firms and nonprofit institutions.[14] Men are not usually involved in day-time work on the household farm, but it restrains many women from taking a job with a firm or nonprofit institution. (The household agricultural work usually adds to other obligations like mother's and housewife's work, and all these together make the woman stay at home.)

A much smaller, yet quite important stratum was that of independent artisans and retailers. Their number has decreased everywhere, to a different extent in each country. Where their activity has remained, it is usually the demand of the population that puts pressure on the authorities to maintain or extend it.

The sphere of the so-called self-employed intellectuals has shrunk considerably; a number of activities carried on individually in the past are now done within firms or nonprofit institutions. Yet there are professions that are almost impossible to conduct in large organizations.

In the case of all the above groups it is the *character of the work*, or a strong social demand for some special activity that limits the absorption effect of firms and nonprofit institutions.

(b) The other large labor resource is *women working in the household*.[15] It is a characteristic feature of a resource-constrained labor market that the proportion of women occupied only in the house is very low among women able to work. Their flow towards employment in firms or nonprofit institutions is often connected with the phenomenon discussed in item (a). When the family-owned farm, or the artisan's and the retailer's shop ceases to function, the remaining housewife's and mother's activities often provide insufficient fulfillment.

This flow and the concomitant change in lifestyle is not complete. What is more, it is in this sphere that a return to the former state of affairs

[14]Small-scale agricultural activity performed by agricultural cooperative members and state farm workers together with their families on small privately owned plots of land is referred to as household farming.

[15]See Kutas–Mausecz (1976).

frequently occurs, within an absorption process which is otherwise quite irreversible. A woman's decision to take a job depends on the wages of the other income-earning members of the family, on the family size, its age composition, state of health; besides, it is not only based on their own views, but also on those of the husband and other family members. The extent of state allowances to mothers staying at home to care for their children may have important effects. In any case, it is quite inconceivable that every woman able to work should take a job in the sector of firms and nonprofit institutions; a considerable proportion of them will remain outside this sector.

(c) In the capitalist system, and even in the transition period, a considerable number of people live *exclusively on profit*, *interest*, *land rent*, *or on selling their property*, although they are able to work. In the socialist system this way of making a living ceases; members of this stratum, if they did not emigrate, generally took jobs in firms or nonprofit institutions.

(d) For the sake of completeness, we also mention *domestic servants*. (In developing countries they are still rather numerous.) Most of these have also been attracted to the sector of firms and nonprofit institutions. At the same time there is a retarding force: certain families which demand domestic servants and who also have the means to pay for them.

Of the above four labor resources it is obviously the first two that are really important; in comparison with them the third and fourth are insignificant.

The list provided above throws some light on the interpretation of the "tolerance limit" as a stochastic phenomenon. It may be that a person X.Y. thought that he would take a job yesterday but changed his mind today, while person M.N. changed his mind to the contrary. As regards the labor force as a whole, we should consider a *tolerance interval*, and the value $\hat{h}$ should be considered as the mean value of the interval. The tolerance limit $\hat{h}$ – as well as other tolerance limits discussed in earlier chapters – is a *social* phenomenon. It becomes effective when certain groups of society *resist* a tendency, in the present case the labor-absorbing effect of firms and of nonprofit institutions. For these groups employment above the tolerance limit is *unacceptable*. But, since the question concerns a regularity of human behavior, this limit is not valid for ever. The tolerance limits of employment may shift in both directions. At any instant, however, the tolerance limit interval is *given*. It is in this historically relative sense that the statement made at the beginning of the section holds true: in the pure resource-constrained system the mobilization of the population able to work is complete, because it has reached the prevailing tolerance limits.

After reaching this limit the participation rate is *more or less stabilized.*[16] From that point it can move only a little, since the tolerance limit can be shifted to a considerable extent only by slow historical processes. Drawing a comparison between the two "pure" systems *the normal participation rate of the resource-constrained labor market is much higher than that of the demand-constrained one.*

There may be certain actions that will shift the tolerance limit, for example if a small town or a village receives a new kindergarten, many of the mothers living in the neighborhood will be more willing to take a job. The *pure* resource-constrained labor market is, however, *defined* so that such shifts can be but marginal. As long as there are, in a real economy, not just marginal but large possibilities for labor mobilization, it has not yet reached the stage of the pure resource-constrained state, but may be, at the most, on the way towards it.

The number of people employable by the sector of firms and nonprofit institutions is, of course, substantially influenced by demographic processes, first of all by the size of $L(t)$, the number of people able to work. In examining the *qualitative* features of the different systems we can, however, disregard this. That is one of the reasons why we describe the characteristics of the systems by relative indicators. Whether $L(t)$ is stationary in time, or decreasing, or increasing, if the labor market in a country has become a pure resource-constrained market the participation rate will approach the tolerance limit and $h^* \sim \hat{h}$.

Thus, just as for the pure demand-constrained labor market, *the normal participation rate is the main characteristic property of the labor market situation in the long run.* It generates the main trend around which instantaneous fluctuations may take place.

(2) Figure 11.3 is analogous to fig. 11.1: the diagram $(z, \boldsymbol{q}, w)$ is now applied to the pure resource-constrained labor market. Interpretation of the axes is identical with that of fig. 11.1.

The point in the upper left-hand corner of the figure represents the normal state of the system. Labor shortage is high, as will be treated in detail in the following section. There is friction in the allocation of labor $(w^* > 0)$, so that the situation where everyone intending to take a job always finds one straightaway cannot occur. The following interrelation is

[16]In Hungary it has been recognized for several years that the available potential reserve labor is virtually exhausted. Using the terminology of Hungarian literature on labor: employment of women has reached the "social maximum", and that of men the "demographic maximum". See, for example, Iván (1975), Karakas (1976), and J. Timár [1977a, b].

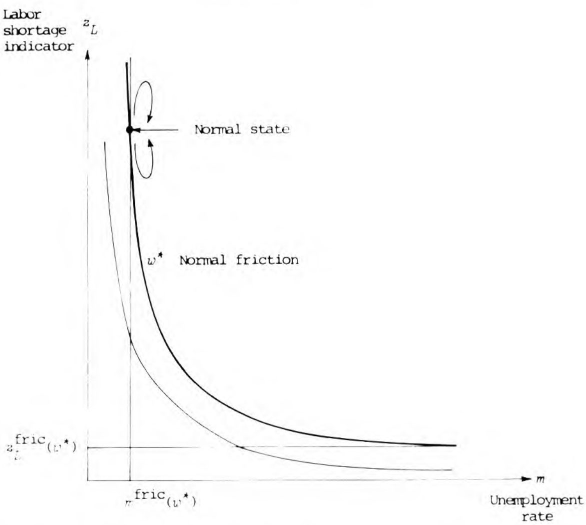


Figure 11.3. Normal state of the resource-constrained labor market.

characteristic of the situation:

$$m(t) \sim m^{\mathrm{fric}}(w^*) > 0, \quad \text{for every } t. \tag{11.7}$$

Unemployment cannot fall much below the minimum level caused by friction, but it does not rise far above it, either. Even with instantaneous fluctuations of employment, it hardly varies around the minimum level caused by friction. Chronic unemployment, the second criterion of the pure demand-constrained labor market, has been totally eliminated.

To this we can add that since (i) there is no potential reserve labor, and (ii) there is no chronic unemployment, frictional "unemployment" clearly becomes an *organizational* problem. True, it is inconvenient for the person involved if he cannot find an adequate job for himself for some reason. (Although in most cases he "searches" while still in his old job, for security's sake.) This phenomenon is not accompanied by the fear of *genuine* unemployment. That is why it is justified to put the word "unemployment" in quotes in this context.[17]

[17]This is in contrast to the pure demand-constrained labor market in which the person without a job is not at all reassured by the fact that he is "only" frictionally unemployed. For the person involved there is no borderline between frictional and "real" unemployment; in any case he feels threatened. In practice, the two phenomena can be separated only with regard to the aggregate.

The iso-friction curve associated with normal friction is itself *given* at a certain moment. In the course of time, however, it may shift; with a series of deliberate actions its shift can be promoted. In fig. 11.3 a second thin curve is drawn, below the currently valid curve (drawn as a thick line). Let us examine the main factors that determine the location of the curve.

(a) The task begins with the medium- and long-term planning of production and labor: the more the composition of labor demand of firms and nonprofit institutions corresponds to the composition of labor supply in regard to professional training, regional availability, and so on, the less friction there will be.[18]

(b) Wage policy plays an important role, both in the long-term wage proportions, and their instantaneous differentiation. This will be discussed in a later chapter.

(c) Labor mobility is closely linked to housing conditions. Housing shortage "binds to the soil". Public transport, infrastructural and social–cultural conditions also play a role.

(d) Changes in the structure of production, and particularly, unforeseen adjustments may necessitate the retraining of labor. The better organized the retraining the less friction appears.

(e) Finally, information also plays a role: whether firms and nonprofit institutions can quickly find those wanting to take the job and, conversely, whether those looking for a job can quickly find someone offering adequate employment.

In fig. 11.3 (as also in fig. 11.1 representing the demand-constrained situation) loops represent cyclical fluctuations around the normal state. In section 9.10 the phenomenon of investment cycles was discussed in detail. This is accompanied – even though in a damped form – by a cyclical fluctuation on the labor market. Labor shortage becomes most intense in the upswing of the cycle and particularly around its peak. At such times hidden labor slack is more fully used with more overtime, or high prizes offered for extra performance.

(3) It was stressed as the third property of the pure demand-constrained labor market that *those used to employment* feel *particularly* defenceless as a consequence of the irreversibility of their position. Now, in the pure resource-constrained labor market, it turns out that although irreversibility also holds there, there is no corresponding defencelessness created by the threat of unemployment since the state of the labor market has changed permanently. The person used to employment has no unemployed competitors in the market, nor is there any possible competition from a huge

[18]See Jánossy (1966).

potential reserve labor. *The behavior of the group used to employment is characterized by guaranteed employment.*[19] This leads us, however, to our next subject, labor shortage.

11.5. Labor shortage

With normal friction $\boldsymbol{w} > \boldsymbol{0}$ *and with a given tolerance limit* $\hat{h}$ *on the normal participation rate there is no further potential external*[20] *labor slack. Therefore, once the pure resource-constrained labor market has established itself, current activity and even more the expansion of the economic system as a whole meets labor constraints.* This becomes one of the most important resource-constraints of the system, basically restricting its growth.

The word "shortage" is to be understood as in other chapters of the book. It is a *stochastic* category, with firms and nonprofit institutions hitting accidentally, at the submicro-level, the thousands of "tiny" labor constraints for each day, that is the many kinds of labor bottleneck. The numerous random occurrences of hitting these constraints have certain probability distributions which might be characterized *statistically*. Labor shortage cannot be described by a single macro aggregate, for aggregate excess labor demand cannot be defined by saying that in total, so many additional staff are needed by firms and nonprofit institutions. This is, like other shortage phenomena, a *vector category*; the intensity of labor shortage must be expressed by a collection of several indicators. From the shortage indicators which may be considered we mention a few examples.

(a) *Initial labor demand* can be defined analogously to initial material demand. (See Chapters 2, 3, and 5.) Its starting-point is the labor demand of the firm for the coming period, assuming that it intends to fulfill the plan, and with the technology specified in the plan. To this demand – derived from the output target and determined by the planned technology – is added, just as for material demand, the surplus demand generated by the hoarding tendency. ("Hoarding" will soon be reverted to.) Let us denote by $\boldsymbol{d}_{\mathrm{L}}^{\mathrm{init}}(t)$ the initial labor demand thus obtained. (The components refer to different skills, etc.) One indicator of shortage is the

[19]Other factors may have an impact in making the worker still feel some dependence upon his superiors. This will be treated in the next section.

[20]As a reminder: in Chapter 2 we referred to the instantaneously unused part of a resource already inside the firm as *internal* slack. Therefore, the attribute "external" means at this point the population able to work, not already employed in the sector of firms and nonprofit institutions, but employ*able* there.

extent to which the number actually employed, $\bar{n}(t)$, falls short of this:

$$z_{\mathrm{L}}^{\mathrm{init}}(t)=\begin{cases} \boldsymbol{d}_{\mathrm{L}}^{\mathrm{init}}(t)-\bar{\boldsymbol{n}}(t), & \text{if } \boldsymbol{d}_{\mathrm{L}}^{\mathrm{init}}(t)-\bar{\boldsymbol{n}}(t)>\boldsymbol{0}, \\ 0, & \text{if } \boldsymbol{d}_{\mathrm{L}}^{\mathrm{init}}(t)-\bar{\boldsymbol{n}}(t) \leqq \boldsymbol{0}. \end{cases} \tag{11.8}$$

(b) In the course of fulfilling the plan, instantaneous adjustment, for example forced substitution, takes place. (See Chapter 2.) Let us assume that part of the labor required to realize the prescribed input–output combination is momentarily not available. The possible reactions of the factory are as follows.

(i) Partial standstill. Because some workers are not available, other workers engaged in strictly complementary activities are held up.

(ii) Improvised modification of the output composition, adjusting it to the instantaneously available labor.

(iii) Making up for labor in short supply by cheaper labor (likely consequence: smaller quantity, poorer quality), or by more expensive labor (likely consequence: wage cost rises and shortages of qualified labor in other work processes).

Examples of all these phenomena may be observed; their frequency and extent can be measured.

It is very important to understand – here as well as earlier in the case of material inputs – that shortage is not *only* a matter of lagging behind demand. Input–output combinations are forced to adjust to it, but the consequences of this forced adjustment also belong to the "shortage" phenomenon.

What is more, labor shortage has a cumulative, spill-over effect. A few workers are missing from the screw factory and therefore output will be less today. In the machine factory the indirect effect already appears in the form of *material* shortage. The work had to stop because there were no screws for the machine.

As a consequence of intensive labor shortage a *sellers' market* predominates in the allocation of labor. Relative power on the market is again unbalanced; the situation is *asymmetrical*, but now, however, it favors the opposite side, as compared to the pure demand-constrained labor market. As a consequence, most of the costly activity associated with allocation are now borne by the buyer. (See Chapter 4.) The buyer, that is the firm or nonprofit institution, does most of the *search*, most of the *information collection*, and it has to *wait* if its instantaneous demand cannot be satisfied.

There are important factors exerting their effect in the opposite direction. Many workers are reluctant to change their job, even if it is easy to find a new one. Others have difficulty in moving because their special knowledge can only be utilized adequately at the given working place, or because they received accommodation from the firm, and so on. Administrative prescriptions may also make it difficult for the worker to change his job at his own initiative. These countertendencies, however, at the most weaken but do not overrule the basic tendency of an asymmetry in the labor market caused by chronic labor shortage.

Along with the reversal of the asymmetry, the main cause–effect relations asserting themselves in the long run within the network of interactions are also reversed. Let us compare formula (11.6) concerning the pure demand-constrained labor market with the following:

$$\boxed{\begin{array}{c}H\\ \text{Population}\\ \text{used to}\\ \text{employment}\end{array}} \rightarrow \boxed{\begin{array}{c}S\\ \text{Labor}\\ \text{supply}\end{array}} \rightarrow \boxed{\begin{array}{c}N\\ \text{Employment}\end{array}} \rightarrow \boxed{\begin{array}{c}D\\ \text{Labor demand}\\ \text{(initial and revised)}\end{array}}. \qquad (11.9)$$

The point of departure is the population used to employment which reaches – in the pure resource-constrained state – its own upper limit and has more or less stabilized. That is basically what determines labor supply. Supply restricts employment. Labor demand – its various phases of initial demand and its revisions in the course of forced adjustments – is basically derived from the former. It is true that at the submicro-level everybody tries to get more labor than is available in total. Yet not even the initial demand breaks away completely from the expected supply, since the plan from which it is derived is based on a forecast of the employment constraints.

11.6. "Unemployment on the job"

While there is no potential external labor slack, and a chronic intense labor shortage is felt, an *internal labor slack exists*. It is also called "unemployment on the job". This was already mentioned in Chapter 2 in discussing all kinds of input, but now the phenomenon must be analyzed separately in regard to labor.[21]

[21]The phenomenon is also well known in capitalist economies. See, for example, Okun (1962). On Hungarian experience of "unemployment on the job" see, for example, J. Timár (1977a, b) and Pongrácz (1976).

It is not enough to observe again that shortage *and* "surplus" may exist simultaneously, for there is also a causal linkage between them. *The more frequent and intensive the labor shortage, the greater will be the internal slack, namely the unemployment on the job.* This relationship will be demonstrated with the example of production in a firm, although it holds – with appropriate modifications – for nonprofit institutions as well.

(1) A considerable part of "unemployment on the job" creates a *non-mobilizable* slack at any given moment, since the complementary inputs required for its utilization are not available. A few of the workers cannot work because their colleagues, indispensable for the collective activity, have not come to work. Or the material has not arrived, parts are missing, the machine has broken down, and so on. (See Chapter 2.)

(2) *Chronic and intensive labor shortage loosens workshop discipline, deteriorates work quality, lessens workers' diligence.* This is not to say that in the resource-constrained state of the labor market there is something wrong with every worker's discipline and diligence. Discipline, care, and diligence are stochastic phenomena. Their degree varies between individuals and fluctuates over time even for a given person. Many factors affect discipline, care, and diligence, but primarily the fact that most people, for the greater part of their working lives already do their work reasonably well without external pressure to do so. And the more they understand the social importance of their work, the truer is this statement. Education, and adequate financial and moral acknowledgement may even further strengthen this effect. We claim only that the *factors operating in favor of discipline, diligence, and care are counteracted by chronic labor shortage.* The worker's *absolute* security, the unconditional guarantee of employment encourages irresponsibility in anyone susceptible to it. Production managers, from foreman to director, may try to impose discipline, but their means of doing so are restricted. The real asymmetry in the labor market (and not their "liberal" disposition) forces them to be indulgent. Even if they decide, as a final step, to sack an undisciplined worker who is disrupting production, he will immediately be taken on for another job, and may even get higher wages.

All this does not lead to the conclusion that managers of firms and nonprofit institutions in the socialist economy are unable to cope with the lack of discipline and responsibility. It only means that the *circumstances conducive to such behavior constantly recur – since labor shortage is necessarily chronic.*

The causal chain now operates in the following way. If labor shortage is more intensive, workers unexpectedly leave their firm or nonprofit institution more frequently, and their places remain unfilled because of shortage.

Alternatively, they may not leave, but simply be absent without justification, or they come to work, but instead of working properly just waste time. In all three cases – as indicated before – the probability that workers engaged on complementary activities will also be unable to occupy their worktime increases.[22] This clearly indicates that the burdens of labor shortage are not only borne by society as a whole, through losses caused by production shortfalls or poor quality, and not only by economic managers who have a lot of problems due to forced adjustment. The workers themselves suffer losses, but these losses are not equitably distributed. It is the careless and lazy who make use of the situation of labor shortage and the diligent who suffer the consequences.

(3) In the traditional economic management system the firm is induced by central allocation and rationing of labor to "reserve" labor. It is worth keeping even possibly superfluous labor in the firm because, if it were relinquished the lower number of staff would be prescribed in the next plan as an obligatory quota.

Intensive labor shortage is bound to give rise to the tendency of "reserving" labor – within the framework of either traditional or postreform economic management systems. This is parallel with what was discussed in Chapter 5 concerning material inputs, where it was called the "hoarding tendency". Even if one or other worker is momentarily not needed for a certain job, the manager does not send him away. He prefers to "hoard" him, out of prudence. Sooner or later, at the latest after a considerable expansion of the particular firm or nonprofit institution, he may be needed. The *hoarding tendency increases labor shortage and unemployment on the job at the same time*. It is a vicious circle, since any increase in the intensity of shortage also strengthens the hoarding tendency.

In discussing the simultaneous presence of labor shortage and unemployment on the job we refer briefly to the role of the so-called "second economy" functioning in socialist countries. Numerous types of activity are included in it. Some are legal, from small-scale gardening by employees in the "first economy" (i.e. the sector of firms and nonprofit institutions) to agricultural activities on household farms by those working in urban factories but living in rural areas. There are, however, semilegal or even entirely illegal forms, such as repairs and services carried out in the worker's free time, without official permission or any tax payment, or

[22]To this is added the indirect effect mentioned earlier. Because of the unjustified absence of the screw factory worker a shortage of screws develops. And, as a consequence, the worktime of some workers in the machine factory may also remain unused.

the same activities carried out during working time already paid by the state-owned firm and sometimes by using material stolen from the firm.

It is beyond the scope of the present book to give a detailed account of the "second economy" and to analyze its role. It is clear, however, that its emergence and constant reproduction can largely be explained by *shortage*. The demand that the "first economy" is unable to satisfy creates a ready market for the products and services of the "second economy".

As a matter of fact, we do not get a full and comprehensive picture of the labor situation if we only consider the sector of firms and nonprofit institutions. In fact, labor supply is larger (measured not as the number of workers but in terms of working hours) than that offered to the sector of firms and nonprofit institutions at legal wages and observing other official restrictions. The surplus is employed in the "second economy". On the one hand, this increases the intensity of labor shortage in the "first economy". On the other hand, it reduces the hidden labor slack in the economy. It reduces the spare time of many workers and also provides them with work during part of the time when they should be working for the firm or nonprofit institution employing them.

11.7. Transition to the pure resource-constrained labor market

After discussing both pure cases separately, let us now study the historical process which leads from the demand-constrained labor market to the resource-constrained one. This takes place under different conditions in each socialist country. Here we shall take just a few of the common features – and even those in a sketchy, abstract, "stylized" form.

The process is demonstrated first in fig. 11.4. The horizontal axis represents time. On the vertical axis the share of potential reserve labor in the total population able to work is measured.[23] Using earlier notation, $(1 - h(t))$.

The time before t_0 is the period prior to the socialist transformation in the country. It is not important to know what type of social formation that country had previously, whether capitalist, precapitalist, or some combination. It may be assumed that the economy was growing then, too, and that the participation rate was growing with it. But in our abstract historical

[23]For a correct interpretation of the figure: it is obvious in the whole reasoning of the book so far that both the time series $h(t)$ and the tolerance limit $\hat{h}$ represent stochastic phenomena. In effect we ought to have drawn stripes with fuzzy edges. To simplify the graphical representation, we used lines.

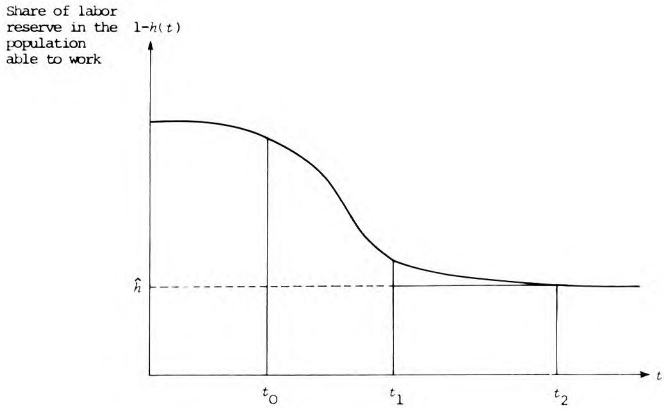


Figure 11.4. Absorption process of reserve labor.

scheme this growth was slow; the normal state of a demand-constrained labor market asserted itself in a more or less pure form.

The historical period $[t_0, t_1]$ saw the absorption of potential reserve labor into the sector of firms and nonprofit institutions. The following historical period $[t_1, t_2]$ represents the *almost-resource-constrained labor market*. Finally, at time t_2, the pure case has established itself, and the labor market is resource-constrained. The absorption of reserve labor has reached the tolerance limit, $\hat{h}$.

It would not be correct to describe the process by a single indicator. Let us use in fig. 11.5 the form of representation used in earlier diagrams, $(\boldsymbol{z}, \boldsymbol{q}, \boldsymbol{w})$. The shortage indicator is not specified; we must visualize the vector describing labor shortage, represented here by only one of its components. As a slack indicator, we do not use unemployment here (unlike figs. 11.1 and 11.3), but an indicator of more fundamental importance, namely the share of potential reserve labor in the population able to work, $(1 - h(t))$. Let us assume – a highly simplifying assumption – that the system shifted along an identical iso-friction curve through the whole historical period. (In reality, surely the curve itself also shifted.)

All through the process *both* labor slack and shortage existed. Both were present in the form of millions of submicro-level stochastic events. During the historical process of change the probability distributions of these random events changed. The former became less frequent and less intensive, and the latter became more frequent and more intensive.

The interpretation of the time t_1 presents a problem. Although there is always – as we have just mentioned – slack *as well as* shortage, the transi-

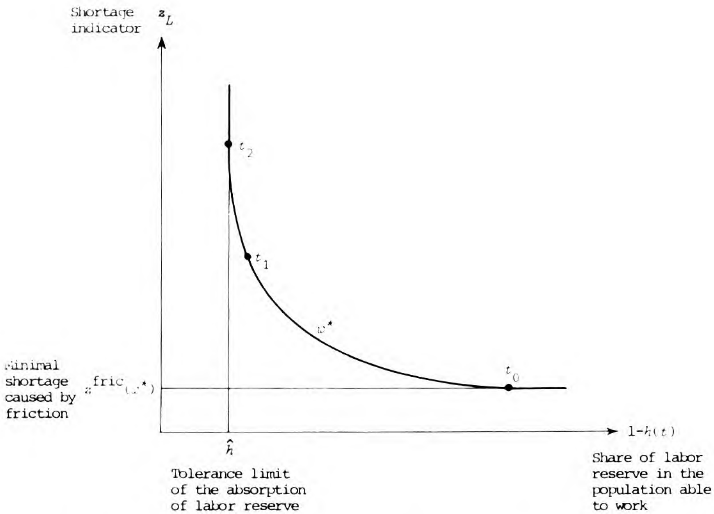


Figure 11.5. Labor shortage and slack in the process of absorption.

tion does not proceed at a uniform rate. There exists a point at which the system, so to say, "tips over" from being predominantly characterized by labor surplus to a predominance of labor shortage. The word "point" is, of course, to be interpreted "allegorically"; in reality the time t_1 can even mean a period of several years. Most important from this point of view are the changes in the *minds* of economic managers; that is, how the labor market situation is perceived and reflected in their decision-making practice. Prior to time t_1, public opinion regards labor supply not as a serious problem; after that time, opinion changes to the effect that it is a serious problem and perhaps even one of the most serious ones. It seems that the recognition of this fact is not a slow diffusion process but is sudden: opinion shifts from one extreme form to the next very quickly. After time t_1 everybody starts to feel labor shortage in their own practical experience; not only labor economists and planners, but every worker employed in a factory, every housewife buying in shops struggling with labor shortage, every patient attended to at a hospital with insufficient auxiliary staff, and so on. Everybody feels as a daily experience what this book calls "hitting a resource-constraint". And, accordingly, *behavioral routines and regularities change, together with different rules of thumb and feedback mechanisms. It is not through a slow continuous movement but more or less as a sudden "tipping over" that the system adjusts to the fact that from now on labor is going to impose an upper limit on various activities.* In this sense the point t_2

is a kind of "gravitation point"; the system is not stabilized in the intermediate state but, once it reached t_1, it will soon start to behave as if it were in t_2.

In the network of interactions between labor demand and supply we described two main kinds of causal direction. One was valid for the pure demand-constrained case (see formula (11.6)), and the other one for the pure resource-constrained case (see formula (11.9)). Around the transition time, we can say that *before "tipping over" the main causal direction was that of "labor demand to labor supply"* (*i.e. formula* (11.6)); *and after "tipping over" it became that of "labor supply to labor demand"* (*i.e. formula* (11.9)).

The first period, $[t_0, t_1]$, is particularly remarkable from this point of view. One of the subsystems, the labor market, is demand-constrained, while all other subsystems (or, we could say, the *whole* of the system) are resource-constrained. The direct *explanation* of the process lies in the nature of the system as a whole. Absorption of potential reserve labor takes place *because* the sector of firms and of nonprofit institutions operates within a resource-constrained system.

The "pure" demand-constrained labor market, that is one of the allocation subsystems, gets stuck in this state – it becomes its permanent, chronic, normal state – if its "environment", that is the other markets or allocation subsystems of the system are permanently demand-constrained. However, in the operation of firms and nonprofit institutions, as soon as the mechanisms that were described in Chapters 2–10 as features of the resource-constrained system or shortage economy start to be effective, absorption of reserve labor *must* begin. *The primary motive force in this is the expansion drive and the closely related almost-insatiable investment hunger*. Growth of the system is held back at every moment by physical constraints and social tolerance limits; it is slowed down by frictions. Yet within these limits the activities and output of the firms and nonprofit institutions expand and, in parallel with that, their labor demand grows irresistibly. *That is why* in period $[t_0, t_1]$ labor demand determines employment, and through this also the process of getting people used to work, and ultimately labor supply. And this continues as long as labor itself does not become one of the effective constraints of the system.

Nowadays, when in advanced capitalist countries the unemployment problem has again become prominent, governments of socialist countries are often "praised" in the Press or at international conferences of experts for their "clever employment policy", which has not deviated from the principle of full employment even amidst capitalist depression. This praise is due, however, not to the *policy* but to the *system*. It is true that much depends on employment policy. (This was discussed earlier in connection with frictions.) Yet an explanation of the process just described depends on considerations deeper than mere government policy. It follows necessarily from the *system*; for if the system is resource-constrained, an almost unlimited investment hunger and expansion drive *must* prevail, and then the potential reserve labor *must* be absorbed sooner or later.

Government economic policy has a considerable influence on *the way* in which this process develops: how long it lasts; what its transitional and persistent features are like; what sacrifices are made by the generation living through the transition period, and so on. Let us consider a few decision problems of primary importance for economic policy, and also the question of their relation to employment.

(1) *Investment rate*. This has a decisive importance from the point of view of the length of the period $[t_0, t_1]$. With a higher investment rate, the absorption of potential reserve labor takes place faster, and in parallel with that, unemployment – the source of much suffering at the beginning of the period – is also eliminated in a shorter time. It seems, however, that a high investment rate is not a necessary condition for the process to reach its conclusion; absorption may also take place at a lower growth rate.[24]

(2) *Composition of investment inputs*. Among the activities involved in fixed capital formation, construction is more labor-intensive, while the production and instalment of machines are less so.[25] It has been observed that the socialist countries carried out highly construction-intensive investments during that period. If they had not done so, the period $[t_0, t_1]$ would have been more prolonged. But the absorption process can take place even with less construction-oriented investment.

(3) *Sectoral structure of expansion*. (*Composition of output*.) The deviation of the period $[t_0, t_1]$ depends on the priority certain branches have in the general expansion of the economy. If labor-intensive branches are expanded most rapidly, the economy will arrive sooner at the full absorption of potential reserve labor.

Which branches have priority, and which lag comparatively behind the preferred branches is of great importance. However, given the question we are asking here, that is concerning the economy's certain arrival at the

[24]Of course, the investment rate cannot be arbitrarily low; the lower limit is related to the growth of population able to work, as well as to the choice of technology. Here we are not in a position to elaborate on the growth-theoretical aspects of the problem.

[25]Imported machines affect employment indirectly: first of all through the labor-intensity of production for exports.

resource-constrained state of the labor market, this is a matter of indifference. This final state will surely be reached by the socialist system, otherwise resource-constrained as a whole, whether it lays stress on heavy industry, or agriculture gets priority. Labor shortages will surely become more severe if infrastructural branches and services are neglected, but the tendency is there even if economic policy starts to eliminate the "postponement" phenomena occurring in infrastructural and service sectors.

While discussing sectoral structure it is also worth mentioning that *not only the total volume of employment, but also the main proportions of its intersectoral allocation are adjusted to labor demand during period* $[t_0, t_1]$. During that period labor flows where there are job opportunities. For example, a new factory is built in an agricultural region. If that is concerned with iron-working, an industrial working class with appropriate skills will develop; if it is a textile or shoe factory, there will be a light industrial working class. *Investment allocation is the leading process; labor allocation follows it and is adjusted to it.*

(4) *Choice of technology.* (*Composition of input.*) There are two opposing tendencies. Engineers together with many economic managers and planners have a prejudice in favor of "modern" techniques, which are usually relatively capital-intensive and less labor-intensive. The other, opposing tendency is that investment resources and particularly foreign exchange allotted for importing machines to provide for the most advanced technologies are very scarce from the outset. As was seen in Chapters 9 and 10, the expansion drive continually hits these constraints. This, however, induces the decision-makers to consider thoroughly how to spend investment resources, and especially foreign exchange. Since there is ample labor in period $[t_0, t_1]$ they are motivated to adopt labor-intensive technologies. Characteristic compromises are made between the two tendencies, usually resulting in a few factories with very up-to-date and capital-intensive technologies, while in other branches, in which development is postponed, the old labor-intensive technology remains. This "compromise" is not derived from calculations based on relative prices, but much more on nonprice signals and nonprice choice criteria within the framework of nonprice adjustment.[26]

It is at this point in the selection of technology where the change from one period to the other is the most manifest, as discussed above for point t_1. Suddenly people start to realize that labor shortage has worsened. The

[26]In Chapter 14 this will be discussed in more detail; at this point only a few ideas have been anticipated.

reaction to this "quantity" signal is to bring labor-releasing technologies and organizational forms into the foreground even in branches that were the victims of postponement in the allocation of investment resources. *The increasingly frequent hitting of labor constraints leads to appropriate short- and long-term adjustments, with a more consistent change of technology.*

In the literature of the socialist countries the following expression is often used: *economic growth has passed over from the extensive period to the intensive one.* At this point we do not intend to give a detailed account of the distinctive features of the two phases, but shall refer only to the difference in employment. In the former, the expansion of production was facilitated by the *extensive* growth of employment; in the latter it must be based upon growth of production per employee, that is on a more *intensive* utilization of labor employed. The intensive period coincides with our period $[t_1, t_2]$ and even more with the period after point t_2.[27]

In the intensive phase of economic growth everybody experiences labor shortage, but there are various theories as to its causes. The most widely discussed views include the following.

(a) Labor shortage exists because production is disorganized and labor discipline is lax.

(b) Labor shortage exists because the prevailing forms of wage control do not stimulate firms to save labor.

(c) Labor shortage exists because the selection of technologies in investments is wrong: it does not adequately provide for the release of labor by more productive machines.

I do not wish to deny the significance of any of those issues, nor the effect they may have on labor productivity. None of them explains, however, the chronic reproduction of *shortage*. Let us make a mental experiment. Let us assume that all three difficulties mentioned above have been successfully eliminated. Production immediately became better organized and more disciplined, and firms are interested in saving labor. A number of new, labor-substituting machines have been installed at the factories. As a result, 10 percent of the total number employed initially might be released. Before much time has passed, the expansionary forces operating in the economy would again absorb this labor into the sector of

[27]It is an important economic–historical and economic–sociological problem to find out what is the chronological order and delay between recognition and reaction. In Hungary a number of economists and planners predicted several years ago the approaching end of the extensive period. (See, for example, Berettyán–J. Timár, 1963; J. Timár, 1964; Fekete, 1973.) Yet the actual drawing of practical conclusions resulting from this recognition and a radical technological change follows only when an increasing number of firms "hit the labor constraint".

firms and nonprofit institutions: it would be employed to increase the volume of output with the existing fixed capital, to expand investments, and to operate the new fixed capital formed by investments.

The "miracle" assumed in the mental experiment accelerated growth in the *real* sphere, which is, of course, a welcome result. Yet shortage is a phenomenon of the *control* sphere of the system, and that would not be eliminated by the "miracle". *Shortage cannot be eliminated by an increase in supply – as long as the inner regularities of the economy make demand almost-insatiable.* Increased supply is also *finite* – while the demand facing it is always driven by insurmountable inner tendencies towards *infinity*.

Labor shortage is part of the more general phenomenon of suction, which we investigated in detail in Chapters 2–5 and later in Chapter 9. The demand of the firm for inputs for current production and expansion is almost-insatiable. While this more general phenomenon persists, labor shortage will always recur.

11.8. "Equilibrium" in the labor market

Finally, we would like to draw a few conclusions from the foregoing. There are a few theoretical methodological ideal which we tried to explain in earlier chapters of the book, but which might be made clearer by the example of the labor market.

In figs. 11.1, 11.3, and 11.5 the Walrasian equilibrium of the labor market corresponds to the point $z = 0$, $q = 0$, $w = 0$; that is there is no labor shortage, no labor surplus, and adjustment occurs perfectly, without friction. This is an abstract reference point highly useful in theoretical analysis, but no labor market has ever been in that state. Real economies are situated somewhere in the interior of the multidimensional space. Any particular system is characterized by the location of the normal state within this space, around which instantaneous fluctuation takes place.

Let us first take the "pure" demand-constrained case. (See fig. 11.1.) What we said about it is compatible with what is called "Keynesian unemployment equilibrium"[28] in standard macroeconomics. For the economist thinking in the conceptual system of modern macroeconomics it is obvious that this is a normal state in the sense that *within the system*

[28]The terminology indicates that this is the problem *described* by Keynes. Historically, however, as we have pointed out, this was characteristic mainly of the *pre*-Keynesian period. As a result of governmental intervention in the Keynesian spirit the state of the advanced capitalist economy deviated to some extent from this "pure" situation.

there operates a control mechanism which – owing to various feedbacks, "traps" and vicious circles – repeatedly returns employment back towards its normal level (i.e. normal unemployment). Under given social conditions this is the "natural gravitation point" of the actual employment level – and by no means the Walrasian equilibrium point. *There is no mechanism existing in reality which would drive the labor market towards the Walrasian point.*

It may make it easier for readers well versed in modern macroeconomics – and possibly other readers, too – to understand the main idea of the *whole* book if we emphasize that this approach *can be generalized.* It can be applied to other markets as well as to nonmarket allocation processes (not only to labor, but also, for example, to the trade in materials, investment resources, consumer goods). It can be applied to other social systems: not only to advanced capitalist countries but also to the socialist system. And not only to one type of asymmetry ("buyers' market", "unemployment"), but also to the other type ("sellers' market", "shortage economy", "taut utilization"). In each case the task is to describe the normal state of the system or subsystem within the indicator space characteristic of that system or subsystem. Afterwards, one has to explain the nature of the control mechanism that drives this system or subsystem back to that normal state following deviations from it. This explanation cannot be reduced solely to a description of government policy measures, but the intrinsic social forces which lead to stabilization around the normal state must be identified.

Control mechanisms and feedback are quite clear in the "pure" demand-constrained case. The *later period of capitalism* is much more problematic. Is it the relatively high employment characteristic of advanced capitalist countries in the 1950s and 1960s that will prove to be the permanent normal state, with the unemployment of recent years being but a temporary cyclical deviation? Or, conversely, was the situation of the 1950s and 1960s a transitory phenomenon (even though it lasted comparatively long), with capitalism now returning to its normal state? From now on, will chronic unemployment and a more or less demand-constrained situation again be characteristic of the functioning of the labor market? Is an advanced capitalist economy able to stop somewhere between the pure demand-constrained and the pure resource-constrained normal states? Is the "fine-tuning" activity of governmental economic policy sufficient to adjust the system to this "medium situation", or are there more deeply rooted internal control mechanisms to do the same thing? Or is it inevitable that internal forces will shift the permanent state of the labor market into one or other definite direction, towards one of the two "pure" cases?

My intention here is just to indicate some questions, which this book does not even attempt to answer. It is a task for other works to achieve a theoretical clarification of these questions. Yet to find a really convincing answer a much longer time – probably the historical experience of further decades – may be needed.

Turning our attention to the socialist economy, more definite assertions may be made because the historical experience is clearer. *The "mixed", "half-and-half" state of the labor market, in which only half of the potential reserve labor has been absorbed, and labor shortage is rather scattered and not severe, will only be transitory. Its own immanent regularities will carry the system towards the "pure" case of the resource-constrained state.*

"Full employment" in the Walrasian sense of the word is impossible. There is no real economic system from which both "excess demand" and "excess supply" have been totally eliminated on the labor market. If, however, we interpret the expression "full employment" in its historical aspect, that is as a full and final liberation from the threat of unemployment, this will become possible. This is guaranteed by the resource-constrained system – but, together with this, it guarantees chronic and severe labor shortage. "Optimization" is not possible: we want full employment, but we do not want labor shortage. They are joint products, which, it seems, necessarily appear together.

CHAPTER 12

Social benefits and costs as a function of social capacity utilization

12.1. Introduction

In Chapter 10, using the example of a road, we found an interrelation between the capacity utilization of a resource (in our example, the capacity of the road being measured by the flow of vehicles on it) and social costs. It was observed that beyond a certain degree of utilization marginal costs grow increasingly steeply. We now *generalize* this interdependence to cover the *whole* of the national economy. Since the example of a road presented in Chapter 10 served as a preliminary introduction to the economy-wide analysis in this chapter, a certain amount of overlapping and repetition is inevitable.

First of all we wish to recall the well-known model of business economics that examines the revenues and expenditures of the profit-maximizing firm (see fig. 12.1). Let us assume that the composition of the firm's output is fixed, so that the volume of production is measurable by a single scalar variable. The theoretical maximum output attainable *from the "engineering" point of view* is called the capacity of the firm. The question of the resource which finally constrains production can be left open; it is not the capacity of one or the other resource that serves to measure capacity, but the measure of output volume. *Capacity is the maximum potential production attainable with given technical possibilities.*[1] Under such conditions the volume of production can be clearly described by the *degree of capacity utilization*, denoted by κ: $0 \leqq \kappa \leqq 1$.

It is assumed that the firm is a price-taker, and therefore its marginal revenue is equal to the selling price of the composite product. This is thus constant, as shown by the horizontal straight line in the figure. The

[1] In reality production never reaches its theoretical capacity, in the engineering sense, if only because in its immediate neighborhood marginal costs rise sharply towards infinity.

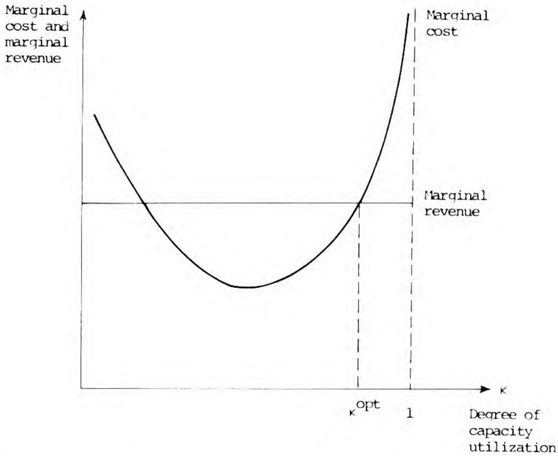


Figure 12.1. Profit-maximizing firm: Marginal cost and revenue.

marginal cost-curve is U-shaped; as a function of utilization it falls for a while, then begins to rise and, approaching full capacity utilization, it rises increasingly steeply.

The growth of marginal costs is counterbalanced – up to a certain degree of utilization – by the increasing revenue of the firm. The profit reaches its maximum at the degree of utilization at which marginal cost and marginal revenue (i.e. the selling price of the composite product) are equal. A higher degree of capacity utilization already begins to reduce profit.

The reason for recalling the model of the firm's U-shaped cost curve in such detail is to awaken associations in the reader. That is because similar interdependencies are also valid in the national economy as a whole. Increasing the utilization of resources not only entails increasing advantages and benefits enjoyed by society, but also increasing costs, sacrifices, and losses borne by society. The U-curve cost model shown in fig. 12.1 is intended only to represent the general nature of the interdependence. Yet we cannot apply the model of the firm to the national economy without some modification to its original form – basically because of aggregation difficulties. In order that – after an appropriate reinterpretation – we can still apply the concepts of "capacity", "costs", and "benefits" for the national economy as a whole, we must indicate precisely the required simplifying assumptions. For the sake of consistency and logical clarity we must nevertheless accept some complexities in constructing the mental experiment.

In the following we shall introduce some *strong* assumptions. These allow us to explain a few characteristics of the social benefit and cost functions. Drawing certain conclusions from these, we shall develop a number of propositions. At the end of the chapter we shall consider whether our propositions would also hold if the strong simplifying assumptions were relaxed or replaced by less restrictive ones. Our initial assumptions are as follows.

(1) We shall conduct the analysis in terms of comparative statics, and in a specific form. The "moment" of analysis is the intertemporal "average moment" of a longer historical period. We could equally well say that the different possible stationary processes of some system are being compared. Within these processes the consequences of certain events appear with regular delays. Whichever moment is chosen, therefore, today's action and the actual consequences of yesterday's action or those of the day before yesterday will be present simultaneously. Our intertemporal "average moment" expresses this as a reaction without any "time-label" to an action also without a "time-label".

For example, the system may be adjusted to a taut utilization of factory capacities, and therefore neglects the maintenance of machines. This will soon have its effects on the condition of the machines. In our "static" mode of expression we would say that an increased degree of utilization has left the machines in worse condition.

(2) All the economic resources of the system are given: the physical national wealth (natural resources, fixed capital, stocks), and the population, with certain intellectual and physical endowments.

(3) The political system is given, together with social relations and all institutions.

(4) The economy is closed; there is no foreign trade. The system can only use its own internal resources.

(5) The output pattern of production is fixed. There are n products in the economy specified in the most detailed disaggregation. Let the unit of product no. 1 (e.g. 1 megawatt-hour of electricity) be the "numéraire". If $\alpha_1 = 1$ unit of product no. 1 is produced, α_2 units of product no. 2 are produced..., α_n units of product no. n are produced – each measured in its own physical units.

The output of product i:

$$x_i = \alpha_i x, \qquad i = 1, \ldots, n, \tag{12.1}$$

in which the variable x is the *output factor* of the system. (This coincides with the volume of product no. 1.)

Our assumption enables us to measure the level of production in the economy without aggregating the different products with the aid of prices.

(6) A fixed proportion of every output is devoted to real capital formation.

(7) Closely related to assumptions (1), (2), and (3), the adjustment properties of the system are considered to be given. Using the terminology introduced in Chapter 8, a given normal friction w^* characterizes the system: the usual degree of imperfect or biased information, the usual vacillation of decision-makers, the usual delays and rigidity in reactions, etc. Let us return for a moment to fig. 8.6. The system is situated on a certain iso-friction curve. In our mental experiment we let the system shift, since we are concerned with *comparative* analysis, comparing several hypothetical states of the system. This shift takes place, however – in accordance with the present assumption – along a given iso-friction curve.

12.2. Utilization of social capacity

With assumptions (1)–(7), it follows logically that total production, and the output factor x representing its level, have an absolute upper limit which it is physically impossible to exceed. This is called the *social capacity* of the system, and is measured on the same scale as the output factor x.

In earlier chapters it was explained that the hitting of physical resource-constraints is a stochastic phenomenon. The economy meets thousands and even hundreds of thousands of bottlenecks at the submicro-level at every moment. As a matter of fact, we might indicate a *capacity interval* for which we could say that with high probability the instantaneous absolute upper limit is situated within it. The upper limit of the capacity interval is the capacity $\hat{x}$ which is considered as given in the framework of the comparative static analysis carried out in the present chapter. Therefore, for the whole society this is a concept analogous to what was called the maximum theoretical capacity of a factory in the *engineering* sense. It is the abstract end-point of a scale and everybody is well aware that in reality it cannot be attained.

Let us denote by κ the *degree of social capacity utilization*:

$$\kappa = x/\hat{x}, \qquad 0 \leqq \kappa \leqq 1. \tag{12.2}$$

The concept of "capacity utilization" was borrowed from business economics, and, as there, the precise magnitude of the capacity $\hat{x}$ has no importance. It is exclusively the *relative* degree of utilization that is interesting in our analysis.[2]

Economics and the everyday language relating to economic phenomena use several concepts that are more or less synonymous with our degree of utilization. Sometimes, for example, "heating" of the economy is mentioned; using this term, "overheating" would be characteristic of a system approaching its capacity interval. We shall not use this expression, however, if only for the reason that it implies a value judgement. In our vocabulary, the variable κ is a descriptive measure; it is a state indicator of the system. If the value of κ is high, that is in itself neither "good" nor "bad". It reflects a certain state, the judgement of which requires separate consideration.

"Employment" is a related category. According to its common meaning the word could be applied in connection with any kind of resource. However, macroeconomics and everyday language have reserved it in practice for a single resource, manpower. Yet in the present chapter we wish to analyze the joint utilization of *all* resources, even though aware of the fundamental importance of labor employment. What is called "Keynesian unemployment" in Western macroeconomics is a *partial phenomenon, one of the components* of a more general and wider phenomenon, which we would now call the low utilization of social capacity.

A further similar category is worth mentioning, namely "tautness" of the economy-wide plan. This was discussed – at least as regards the firm – in section 3.2. The relationship between tautness of plan and hitting resource constraints was explained there. The attributes "loose", "taut", and "over-taut" refer to various regions of the same scale which also measures the degree of social capacity utilization.

12.3. Social benefits and costs: Methodological remarks

We begin by conceptual clarification and a few methodological remarks.

All the phenomena that contribute to the welfare, pleasure and contentedness of individual members or groups of society are called *social*

[2]Later we shall use the expression: "social capacity utilization, κ, comes close to 1", without demonstrating numerically what the word "close" means. It is possible that $\kappa = 0.95$, but perhaps $\kappa = 0.87$, depending on how exactly we determined $\hat{x}$. A system comes close to full utilization of social capacity if the managers of production see with "engineers' eyes" that with the usual frictions of the system it is impossible to obtain substantially more output from the factories.

benefits,[3] and the phenomena that involve burdens, loss, suffering, sacrifice or trouble to individual members or groups of society are called *social costs*. "Social benefits", just like "social costs", do not represent a financial category. Some of the phenomena listed above are usually measured in money terms (e.g. volume of consumption, on the benefit side; or labor input, on the cost side). However, we also consider to belong here such factors as are not usually expressed in money terms (e.g. feelings of security, on the benefit side, and fears, on the cost side).

We wish to refrain from aggregating benefits or costs of various types using weights determined *a priori*. (We shall revert to this in a subsequent section of the chapter.) We should like to avoid bringing *value judgements* into the *description*, but we shall not be entirely consistent in this respect. As a matter of fact, it is already a value judgement to regard one social effect as a "benefit" and another as a "cost". We shall consider social benefits as well as social costs as *vector* categories. Let the former be denoted by vector $\boldsymbol{b}$, and the latter by vector $\boldsymbol{c}$.

Each component of the two vectors, that is each indicator of social benefits and costs, is measured in its own particular units. We recall the example of the road, with petrol consumption in liters, the accident situation described by the frequency of occurrence of events of various degrees of gravity, and so on.

Both vectors have, as a matter of fact, large numbers of components. In our analysis we select from both vectors the different characteristic *groups* of components and examine them closely. Our hypothesis is that the social benefit indicators or social cost indicators associated with one or another such group are functions of the degree of social capacity utilization.

We do not offer an exact description of the functions; our efforts are less ambitious. We shall take in turn each group of social benefit and social cost indicators and argue in support of the *existence* of the functions. In addition we can make a few observations about the *properties* of the functions (whether they are increasing or decreasing; how the derivatives behave, etc.).

In the diagrams, social capacity utilization, the variable κ, is measured on the horizontal axis: $0 \leqq \kappa \leqq 1$.

In graphical representation we shall meet with similar difficulties as before: on two-dimensional paper we cannot represent a multidimensional vector on a single axis. Therefore, each cost group is symbolized by a

[3] We follow the usual interpretation of "social benefit" and "social cost" as applied in welfare economics and cost–benefit analysis.

representative component, and the curves will always represent the first derivative of the benefit or cost function, that is the marginal benefits or costs.

The shape of the curves corresponds to the properties that are described in the text; otherwise, their specific form is arbitrary. This is, however, indifferent, since they serve exclusively to demonstrate the main features of the relationship.

12.4. Social benefits as a function of utilization

It follows from assumption (5) above as well as from the definition of κ that as a function of capacity utilization, the production of all n kinds of products grows proportionately. Since with assumption (6) the investment ratios were fixed, the *material consumption of the population* grows in parallel with production.

We shall not consider whether it is "good" for society if material goods are available in a larger quantity. Hungary has emerged from poverty not so long ago, so for a Hungarian economist the answer is obviously "yes". Growth of production and consumption are recorded on the "benefit" side of the consequences of a higher degree of utilization.

According to assumption (6) a fixed share of production is used for investment. Therefore, a higher degree of utilization κ permits *faster growth*, at least of real capital formation.[4] If a larger volume of material goods is regarded as a "benefit", a faster growth rate of their production and consumption must also be considered, *ceteris paribus*, as a benefit. Although we are conducting a static analysis, this dynamic effect of a higher degree of utilization κ can also be recorded as benefit.

As a function of social capacity utilization, *the employment of labor grows*. Although employment was the subject of an earlier chapter, we shall soon touch on it in the analysis of costs, and briefly comment on it here. We need not say that higher employment means more production, since that would be a double counting on the benefit side. (We have just said, a few lines above, that an increase in κ means more production.) What has to be added to this *material* achievement is the *moral* effect of the extent of employment. The unemployed are not only deprived of wages (this might be partly or fully compensated by unemployment dole), but they also suffer from the humiliation of idleness. Full employment not only in-

[4]This would only be a *sufficient* condition for faster growth with Harrod–Domar-like assumptions.

creases wages, but also strengthens security, and human dignity. It establishes the material foundations for equal rights for women. That is why we are justified in recording it as a separate item among the benefits that are increasing functions of capacity utilization.

12.5. Internal physical inputs for production

Let us now go on to examine social costs. Their first group is the *internal physical inputs for production*. Those inputs – materials, labor, wear-and-tear of machines – come into this category which would be regarded as production costs in either a capitalist or a socialist firm, and would be accounted, self-evidently, in money terms. Although the behavior of these costs is fully treated in books on business economics, repeating it will do no harm. First, let us consider a single firm; "capacity" means, for the moment, the capacity of the firm in the usual interpretation of business economics. Four types of costs are usually distinguished, according to their dependence on capacity utilization. The types are presented – with one exception – in fig. 12.2.

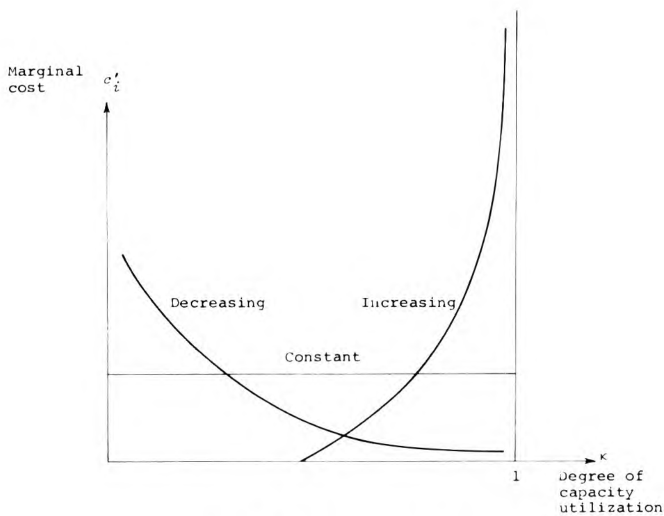


Figure 12.2. Internal costs.

"Fixed" cost, not changing as capacity utilization varies. A major proportion of administrative outlays belongs in this category, as well as the maintenance of buildings, and so on.

Marginal cost a decreasing function of capacity utilization. This includes, for example, human and mechanical "stoppages". Even though the worker gets his wages, it is certainly a loss for society if he is at his working place but cannot do his job. The machine loses value, and its physical condition may be deteriorating even while not used.

Marginal cost a constant function of capacity utilization; in other words, the proportional costs of production. Most prime costs come into this category, such as labor inputs, direct material inputs, and so on.

Marginal cost an increasing function of capacity utilization. We turn to this case with special interest, since we focus on utilization levels approaching full capacity, and it is here that increasing marginal costs appear. A few examples may be given.

(1) Night shift and work done on holidays are special burdens on the worker. The more the firm wishes to get out of its given capacity, the more important becomes night shift and work done on holidays. This is a burden, by the way, not just on the worker and his family, but it causes other losses as well, for productivity is lower, and quality usually poorer in shifts accomplished outside regular working hours.

(2) Approaching resource constraints, forced substitution becomes more frequent. This was discussed in more detail in Chapter 2. The tauter the production plan, the more frequently one or another input is not available. Attempts are made at replacing inputs in short supply, and more expensive or cheaper inputs are used, leading to social losses.

(3) The larger the output a firm wishes to achieve with the resources instantaneously available, the more it neglects some of its permanent tasks, such as the maintenance of machines, product development, etc. In seeking the quantitative growth of production, firms do not adequately improve the quality of products, and may even allow it to deteriorate. Losses suffered as a result contribute to the increasing marginal costs.

If we measured the four types of cost in terms of the appropriate prices, and aggregated them, we would arrive at the *U-shaped marginal cost function*, already presented in fig. 12.1. As regards the first group of social costs, the internal physical inputs for production, we do not doubt the possibility of aggregation using prices, since that is actually done in both capitalist and socialist firms. Yet, aside from aggregation problems, it is sufficient for the present argument to emphasize the following. *Within the firm, both decreasing and increasing marginal costs arise as a function of* *capacity utilization. The more we approach the capacity limit, the steeper the increase in marginal costs becomes.*

And, if all this holds for one firm, it must also hold for the aggregate of firms. Therefore, within the first group of social costs, the sphere of internal physical inputs for production, both the phenomena of "increasing returns" (decreasing marginal cost), and that of "diminishing returns" (increasing marginal costs) also appear at the level of the economy as a whole.

What we said above about internal physical inputs in the sector of *firms* also holds for the most part for the internal physical inputs of *nonprofit institutions*. Because of the limits of the present book, we shall not go into a more detailed analysis of these interdependencies.

12.6. External physical losses and burdens

The next group of social costs is also closely related to the activities of firms (and partly also of nonprofit-making institutions). But these do not appear directly in the accounts of the firms and nonprofit institutions as costs accounted in money terms. In microeconomics they are called *external effects*. We mention a few typical examples.

(a) With increasing capacity utilization the probability of some deterioration of the workers' health increases, as does the likelihood of accidents.

(b) Similarly, damage to the natural environment, including air, water, flora and fauna, is likely to increase.

Both types of effect can be eliminated, or diminished considerably, for example by increased labor safety, prevention of accidents, the introduction of antipollution equipment, and so on, but these all require very large inputs. Either as the original problem, or as the inputs to eliminate it, the effects mentioned in (a) and (b) always give rise to social costs.

Today both socialist and capitalist firms incur some of these "preventive" outlays themselves. Thus, previously external costs become at least partly internal. In this way, the borderline between the first (internal) and second (external) groups of costs may change historically.

Effects (a) and (b) are very widely discussed in the literature; therefore, important as they are, we need not discuss them any further.

(c) At this point it is worth mentioning the phenomenon called "overcrowding" or "congestion". In referring to it, we go beyond "production" in the narrow sense, that is the production of material goods, and must also take into consideration material and nonmaterial services, as well as every other kind of human activity.

One element of what we have called social capacity is *space*: natural space and man-made space. If κ is increasing and approaching 1, that is the capacity limit, it will also entail that in a given space – natural and man-made – more and more human activity takes place. The available space becomes increasingly crowded.

(i) Industry, urbanization, governmental administration, and the development of large cultural and trade centers confine "green" nature into an ever narrower space.

(ii) In a given urban area the growing population, and the increasing level of activity, make the town "dense": the streets, squares, and public buildings become congested.

(iii) The communications network becomes increasingly congested. We referred to it in connection with the road example. But the problem arises more widely, including with public transport vehicles, airports, car service, and so on.

(iv) More activity with given capacity entails the cramming of more machines into the same workshop, more desks or drawing tables into the same office, more beds into the same hospital-ward, more desks into the same classroom, and so on.

Congestion, overcrowding, and "bustle" have harmful effects on people's general condition and social life, as will soon be discussed. At this point we emphasize the "physical" effects, including damage to health, and accidents. But it is also detrimental to the quality of work: the worker will more often make a faulty product, and the designer a faulty drawing.

(d) Increase in employment is concomitant with the extremely important social benefits mentioned in section 12.4. We must not forget, however, that the same process also involves high indirect social costs. We shall discuss only one example, a few effects of the employment of women. If a woman who has been working in the household takes a job with a firm or nonprofit institution, some of her previous activities will be carried out by other institutions, such as nurseries, kindergartens, school day-centers, cleaners, the catering industry, and so on. Some of the work will also be done by her husband, prolonging his worktime. And finally, her own working hours will also be extended, with the first shift at her work place followed by the second shift at home.

Summarizing the economic features of the second group of social costs, the external costs (a)–(d), we can say that they belong to the category with *increasing* marginal costs. Approaching the upper bounds on social capacity they also increase more steeply.

12.7. Maintenance of the government machine

The third large group of social costs to be considered are the *outlays on the government machine*. Let us see how these costs are related to κ, social capacity utilization.

These outlays are mainly fixed. We saw with the first group of costs that outlays on the firm's central administration are largely independent of the firm's output. A similar observation can be made about society as a whole, with a considerable proportion of central government's administrative inputs remaining unchanged, if production increased or decreased by 20 percent.

Several types of governmental expenditure grow with production, not proportionately but at a slower rate. Maintenance of the highest organs directly regulating the economic sphere fall into this category.

All this implies that as a function of social capacity utilization, the state apparatus may become relatively "cheaper". Maintenance of the state is a "social overhead cost". If production grows, the overhead cost per unit of production will be reduced.

An opposite effect also arises; as κ approaches 1, and the economic situation becomes more strained, it can be expected that administrative–bureaucratic activities will grow. The organization of material allocation will become more complicated, an increasing quantity of goods and services being drawn into the system of administrative rationing, and so on. Complaints and discontent become more frequent, inducing the state to be alert and intervene more frequently. The hypothesis that, as a function of social capacity utilization – in the higher region of utilization – increasing marginal costs also appear in the field of governmental expenditures, seems to be justified. Empirical testing of the hypothesis requires some research by political sociologists.

12.8. Public feeling

In the preceding three groups of social costs, *physical* inputs and losses incurred as a function of utilization were discussed. In the fourth group their "mental" aspects will be treated. People's *disposition, contentedness, and general feeling* react to a number of economic phenomena closely connected with the utilization of social capacity. The question concerns an intricate network of interrelations, appearing in a variety of forms in different social systems and, within a given system, in the different classes and groups of the population. We only consider two issues related to these networks.

The first appears at a low level of social capacity utilization, and is always associated with mass unemployment. Aside from the economic losses caused by this to those directly affected as well as to the whole of society, it also engenders various social tensions. It is a syndrome with many different symptoms, including criminality, acts of terrorism, suicides.[5] If social capacity utilization increases, and employment grows along with it, those negative phenomena will become rarer. In this sense, therefore, the effect belongs to the category with *decreasing* marginal costs. This is clear both in the capitalist economy when the employment situation improves, and in the early historical period of the socialist system, when absorption of potential reserve labor and the unemployed begins.

As was explained in Chapter 11, after a certain transitional period, the socialist economy arrives at full employment, a result *guaranteed* by the inner laws of movement of the system. From that point – that is above a certain degree of utilization of social capacity – the social effects appear not only on the cost side, in the reduction of the disadvantageous consequences of low employment, but also on the *benefit* side.

Secondly, let us look at the other end of the κ-axis, corresponding to a high degree of social capacity utilization. Some effects were discussed in earlier chapters, while others will be treated in Part II.

(1) The closer we come to full utilization, the more frequently we experience bottlenecks in production, as well as in other fields of human activity. This not only causes physical loss, but also creates a tense atmosphere. Shortage, forced substitution, improvisation, and rush make people impatient, overhasty, and irritated at work, affecting director and foreman, worker and clerk.

(2) It is not only as a producer, that is a supplier of goods or services, but also as a buyer that one gathers such an impression. The purchasing agent, the engineer implementing investments, and the housewife all share this experience. Being at the seller's mercy, queuing, waiting, searching for

[5]Here we return to the phenomenon already mentioned in section 12.4. There, in discussing the benefit side, we pointed out the atmosphere-*improving* effect of full employment. And now, in discussing the cost side, we mention the atmosphere-*deteriorating* effect of unemployment.

goods, forced substitution performed in the course of purchasing – all give the buyer many bitter hours.[6]

(3) Taut production may entail longer worktime, which also affects morale.

(4) Investment tension might lead to investment absorbing some resources intended for consumption. This may slow down the rise in living standards, and may cause difficulties in one or other field of consumer goods' supply. It may influence the general mood of the population.

These effects must be listed in the group with marginal social costs *increasing* as a function of utilization. The closer we come to the capacity limit, the more sharply they increase. We have now used the apparatus of the present chapter to describe a phenomenon that we have already treated (last time in Chapters 9 and 10). The "strongly felt deterioration of social atmosphere" as a specific kind of steeply rising marginal "social cost", and the "transgression of social tolerance limits" – these are *synonymous* expressions. We cannot determine which are the narrower constraints: the *acceptance* constraints of society, that is the tolerance limits, or the *physical* constraints on production, that is the resource-constraints. This also depends on the particular political conditions of the system. There is a close interaction between the two kinds of constraint, in any case. It is felt that society is close to its tolerance limits precisely when the hitting of resource-constraints becomes increasingly frequent and troublesome.

We do not put an equality sign between the decreasing marginal costs discussed in the first half of this section, and the increasing marginal costs discussed in the second half; the less so, since such an "equality sign" would be entirely alien to the whole spirit of the book. We have stressed several times that it is theoretically impossible to compare one kind of "good" or "bad" to another kind of "good" or "bad". In the case of these two specific kinds of costs we can add something more. Viewing it in a historical perspective it is certainly true that the elimination of unemployment is incomparably more important than the disadvantages accompanying the worsening of shortage. But it is also true that people do not think in a historical perspective at all moments of their life. The employee of a firm irritated by material shortage, or the young couple unable to get their own home will not be pacified if reminded that they should be glad about the elimination of unemployment.

[6] A number of shortage indicators discussed earlier are also suitable for the measurement of a few of the social costs, sacrifices, and losses caused by shortage. That is to say, the shortage vector z and the social cost vector c have a number of components in common.

And this brings us to a more general sphere of problems, concerned with the comparison, aggregation, and evaluation of social benefits and costs so far discussed individually.

12.9. "Welfare optimum"

Those believing in a social "welfare function" would be led by the analysis above directly to the "solution" of the problem presented in fig. 12.3. Their train of thought would be the following:

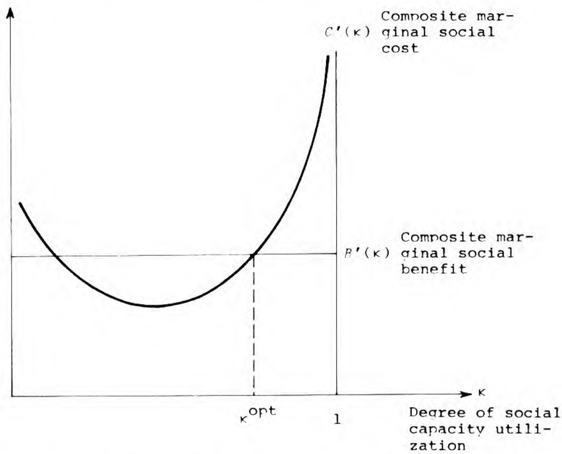


Figure 12.3. "Welfare optimum" in the case of U-shaped marginal cost.

On the one hand the *composite* social benefit function must be determined:

$$\underbrace{B}_{\boxed{\text{composite social benefit}}}(\kappa) = f(\underbrace{\boldsymbol{b}}_{\boxed{\text{vector of social benefit indicators}}}(\underbrace{\kappa}_{\boxed{\text{degree of social capacity utilization}}})). \tag{12.2}$$

In drawing the figure it was assumed, for the sake of simplicity, that B is linear in κ, so that the derivative is a horizontal straight line.

On the other hand, the *composite* social cost function must also be given:

$$\underbrace{C(\kappa)}_{\boxed{\text{composite social cost}}} \quad = \quad \underbrace{\varphi(c(}_{\boxed{\text{vector of social cost indicators}}}\underbrace{\kappa}_{\boxed{\text{degree of social capacity utilization}}})). \tag{12.3}$$

In drawing the figure it was assumed that it is U-shaped: first it has a decreasing, then an increasing section.

Then the *welfare function* is simply

$$\underbrace{W(\kappa)}_{\boxed{\text{welfare function}}} \quad = \quad \underbrace{B(\kappa)}_{\boxed{\text{composite social benefit function}}} \quad - \quad \underbrace{C(\kappa)}_{\boxed{\text{composite social cost function}}}. \tag{12.4}$$

The location of the "welfare optimum", κ^{opt}, is at the second intersection of the two derivatives. In fig. 12.3 the optimum degree of social capacity utilization is well below full utilization.

This is a fine and simple solution, which expresses a trivial yet highly important truth. *One must not always strive at all costs to achieve the maximum utilization of social capacity. It is not worth achieving it if, approaching full utilization of capacity, the marginal social cost already exceeds the marginal social benefit.*

Nothing more profound can be learnt from fig. 12.3. It cannot be considered really as a "solution", either in a descriptive–explanatory theory, or in a theory for elaborating practical recommendations; at most it produces the illusion of a solution to the problem. The assumption that the function B is linear was arbitrary, and it could as well take some other shape. It is open to question how its numerical parameters can be estimated. (For example, how many kilogrammes of meat are equivalent to the sense of security produced by full employment.) The assumption that the marginal composite social cost function C' was U-shaped was also arbitrary. It is only U-shaped if those who form the value judgement expressed by this welfare function recognize and acknowledge the increasing marginal social costs. If they do not take these costs seriously, or even neglect them entirely, it will be exclusively the effects of the constant and increasing returns that will appear. In this case we would obtain a function C' similar to that presented in fig. 12.4.[7] And in that case the welfare

[7] To simplify comparisons it is assumed that B is identical in both figures.

function reaches its maximum at maximum capacity utilization. That is, the "welfare optimum" now sanctions the tautest expansion.

The theoretical determination of the "welfare optimum" does not solve anything, it just translates the formulation of genuine, fundamental problems from one language into another. The *original* dilemma is as follows: the high- or low-degree of utilization of social capacity has various social consequences. What is the relative importance of these consequences, both

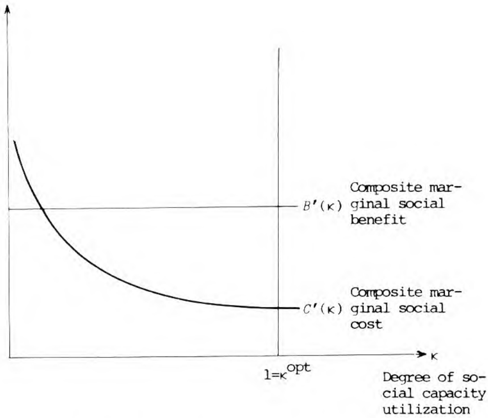


Figure 12.4. "Welfare optimum" with decreasing marginal cost.

in real practice, and according to some explicitly given set of value judgements? The same dilemma *translated* into the language of welfare economics concerns the shape of functions *B* and *C*, and the numerical values of their parameters. The latter "translated" question is no easier to answer than the original one. On our part, therefore, we shall omit this "translating" operation from our analysis as a superfluous stage.

12.10. Observation and measurement tasks

The discussion of the welfare function and the measurement of welfare is carried on at two levels. One is that of pure theory,[8] which we cannot even

[8]See the dispute over Arrow's "impossibility theorem" (Arrow, 1951). The dispute is surveyed in the volumes Arrow-Scitovsky (1969) and Phelps (1973). Furthermore, see the discussion concerning Rawls' ideas about "distributional justice". (Rawls, 1967; Phelps, 1973.)

touch upon here since it would lead us far from the main topic of the book. The other plane is of a more practical character, and centers on the following question: is GDP or any other aggregate output indicator suitable for the measurement of the development and welfare of society?[9] According to one school of economists and statisticians, the answer is *yes*; at the most it should be adjusted, to reflect more exactly the benefits and costs so far unaccounted for in the aggregate output indicator. The other school's answer to the question is *no*; on this view the various aspects of development and welfare, and the performance of the economic system, must be described by *a large ensemble of social indicators*.

It is clear from the reasoning in this book, and particularly from the present chapter, that I support the second school. This is expressed in the idea recurring as a leitmotif of my work, that it is *impossible* to describe the state of processes with inherent *inner contradictions* by a single comprehensive scalar indicator; vectoral measurement is *inevitable*. Scalar measurement adds magnitudes of identical sign but which are incommensurable; besides, it "nets out" magnitudes of opposite sign, the balance of "good" and "bad", thus obscuring the inner contradictions.

With others in the second school mentioned above, I acknowledge that GDP or any other aggregate output indicator may play a useful role in surveying the whole economy, because it reflects in a "condensed" way many different benefits and costs. Trouble begins if a "total" and "all-comprehensive" character is attributed to these aggregate output value indicators. But this danger does exist, and is most closely connected to the phenomena examined in detail here, particularly quantity drive and taut expansion up to physical limits set by resource constraints. If we accept GDP[10] as a "proxy" for the variable W in the welfare function (12.4), it is appropriate to aim for a maximum GDP. And in that case an economic policy must be pursued which drives GDP up to the instantaneous physical constraints.

Vectoral measurement, that is the observation and measurement of a

[9]See, for example, Allardt (1973), Nordhaus–Tobin (1972), and Stone (1975), and in the Hungarian literature Andorka–Illés (1974), Dániel (1977a), Ehrlich (1967, 1968), Hankiss–Manchin (1976), Jánossy (1963), Kornai (1972a, b), and Rimler (1976).

[10]From the aspect of the problem under discussion it is immaterial whether it is GDP or the national income calculated according to the MPS accounting system that figures as the indicator of aggregate output.

large number of social indicators, is no easy task, but seems to be completely solvable. In section 12.3 it was stressed that *theoretically* the vector of social costs (and, similarly, that of benefits) consists of many components. Nobody thinks, however, that *in practice* all should be regularly observed. Statisticians will surely find a subset of indicators whose measurement would not be too expensive, yet which can satisfactorily represent the state of the system. In addition, of course, various "condensations" are necessary: partial aggregations and other synthesizing devices, in order for the information to be sufficiently transparent.[11] The important thing is that the contradictory character of the processes should not be lost in this condensation. The indicators should show that as a function of social capacity utilization, achievements *and* problems, decreasing *and* increasing social marginal costs appear separately.

If in the coming years the regular observation and measurement of social indicators progresses, we shall be able to test *empirically* a number of hypotheses that have been explained in the present chapter about interrelations between $\boldsymbol{b}$, $\boldsymbol{c}$, and κ, that is of social benefits, costs, and the degree of social capacity utilization.

12.11. The planner's attitude and his "conditioned reflexes"

Many issues to do with observation and measurement remain unsolved, which certainly makes it more difficult for planners to see clearly. Yet, however important a more balanced and comprehensive statistical observation may be, it cannot be expected to change essentially the behavioral regularities of an economy. We shall understand the problem more deeply if we try to appreciate the *characteristic attitude* of central planners in the traditional management system, with the "conditioned reflexes" underlying their decisions. These do not originate in any individual's mind, but in their *role*, which determines that in specific social relations they have to decide certain questions.

The problem is illustrated by fig. 12.5. The continuous lines represent

[11]These methodological problems arose – although in another context – already in section 7.9.

marginal costs and the dotted lines marginal benefits. We show a number of curves to indicate that the question is about vectors. The curves of marginal *benefits*, and those of *decreasing* and *constant* marginal costs are thick; this is to suggest that these are carefully observed by those who shape economic policy. The curves of *increasing* marginal costs are thin, suggesting that these are only faintly perceived by them. What causes this particular duality of attitude and perception?[12]

In order to give the discussion a "taste of reality", let us imagine a hypothetical decision situation. The end of the year is approaching, so that the national plan for the coming year must be prepared. Inputs required for the governmental apparatus have already been decided, and the initial resources at the beginning of the new year, as well as social capacity, are practically determined. We can choose between two versions of the plan, the first giving 3 percent more GDP than the second. That is, $\kappa_1/\kappa_2 = 1.03$.

A simplified terminology will be used below. The difference $b_i(\kappa_1) - b_i(\kappa_2)$ is called the *increment of benefit* of type i. The difference between actual costs at capacity utilization κ_1 and the hypothetical cost computed with a linear cost function passing through the point $c_h(\kappa_2)$ is *called the reduction of cost* (or *the increment of cost*) of type h:

$$\text{reduction of cost} = \left(c_h(\kappa_1) - c_h(\kappa_2)\frac{\kappa_1}{\kappa_2} \right)_-,$$

the difference is negative (decreasing marginal cost);

$$\text{increment of cost} = \left(c_h(\kappa_1) - c_h(\kappa_2)\frac{\kappa_1}{\kappa_2} \right)_+, \tag{12.5}$$

the difference is positive (increasing marginal cost).

(1) Increments of benefit and reductions of cost are much more *certain* than increments of cost.

Increments of benefit mean first of all 3 percent more output, which can clearly be put into the numerical targets of the plan. It is also certain that

[12]The "rhythm" of the answer is similar to the train of thought in the answer given in section 10.5 to the following question: why does the decision-maker feel the temptation for postponement? The similarity is explained by the fact that the question asked there constituted *part* of the question now put *more generally*.

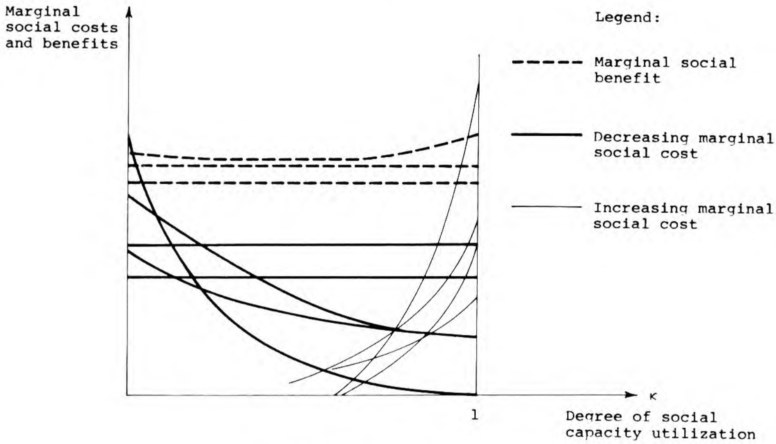


Figure 12.5. "Duality" of perception.

the 3 percent higher production will neither increase the central administration, nor increase administrative costs in nonprofit institutions or firms. Therefore, social "overhead costs" per unit of product will be so much less. All this is "tangible" and *can be planned*. What is on the other side, the side of increasing marginal costs? More forced substitution, more search, more waiting, and so on. But who can be sure that this depends on that particular 3 percent growth in production? And who could say how much loss that would entail? Between the components of marginal social benefit and cost and the degree of utilization of social capacity, the relationship is stochastic. That is so, but for benefits and decreasing marginal costs the correlation is close, while for increasing marginal costs it is rather weak. This is one of the primary explanations for the weaker perception of the "thin lines".

(2) Most of the increments of benefit and reductions of cost are *internal*: their monetary reflection appears in the accounts of the government budget, nonprofit institutions, the firm, and the private household. As opposed to this, most increments of cost are *external*. (This problem refers back to the measurement problems discussed in section 12.9.) In order to illustrate this problem of "internal versus external" we shall move on for a moment to the example of the family household, though this will be

discussed more fully in Part II. If, as a result of the tauter version of the plan, the volume of consumer goods and services for families were to grow by 3 percent, this would be expressed in internal form in the family accounts; in the household and consumption statistics. Accordingly, it is easy to list it among the plan targets. If, in parallel with this, longer queuing is needed; more search has to be done for goods; the seller is even more rude and peremptory with the buyer; the producer and seller are even less interested in quality improvements; the buyer has to accept forced substitution even more frequently – all these are external consequences.[13] They are not accounted for even in the records of a household. They do not appear in household and consumption statistics, therefore they cannot appear separately among the plan targets, either.

We have now cited the example of the household, though a number of similar phenomena could also be mentioned taken from production, trade, or other social activities. All effects of an external character are more difficult for the planner to perceive.

(3) The increments of benefit and reductions of cost can be disposed of centrally. The plan itself can decide on the distribution of the net gains between investment and consumption; the share of various sectors in the gain to investment; and the division of consumption between the various groups in the population. On the other hand, cost increments are *decentralized and atomistic*. To a certain extent they affect every member of society, and in all their roles, as producer, buyer, consumer, user of the health service, tenant, and so on. Yet all these losses are quite small, taken individually and at a single moment of time; because of their scattered character they are thinly "distributed". This is another important explanation for the dual perception – "thick line–thin line" – whereby the hope for increments of social benefit and reductions of social cost has more influence on the decision-maker's conditioned reflexes than the possible danger of increments to cost.

(4) Finally, a special *time preference* also plays a role. Increments of benefit and reductions of cost are apparent *immediately*, while increments of cost only appear *with a lag and extended over a long period*. The taut utilization of resources, "rush-work", may, for example, entail the neglect of product development, quality improvement, machine and building maintenance. All this will certainly bring its own punishment, but only later, slowly and gradually. Public feeling does not react immediately to

[13] Using the apt expression of the book by Robinson and Eatwell (1973): the consumer's purchasing power grows, while his "shopping" power falls.

every irritant either; such phenomena often undermine people's contentment little by little.[14]

We are talking about a special kind of time preference: this is not the "hedonistic" preference that is used in standard consumption theory. It is not for today's "enjoyment" that we renounce today's accumulation, which could yet provide for more consumption, more "enjoyment" tomorrow. The "conditioned reflex" we are talking about prefers *today's output*, and with it today's consumption *plus* today's investment, accepting, as a counterpart, such disadvantages as will harm production, consumption, and investment equally at present and in the more distant future. The "reflex" may be strengthened by political–ideological factors, and by a belief in the uniquely favorable effects of the fastest possible rate of economic growth. *Impatience for the sake of the fastest possible expansion today*, even at the cost of expansion in the more distant future – that is the characteristic feature of the time preference in question.

Knowing all these factors, our forecast of the choice between the two plan variants is clear. If he does his work professionally and conscientiously, the planner will carefully consider whether the tauter version is *feasible*. Using the terminology of the present chapter, he asks whether $\kappa_1 \leqq 1$. If so, he will choose *that version*.

We have listed four factors that strongly influence the planner's attitude, the duality shown in the perception of benefits and costs, and the associated conditioned reflexes affecting decisions. The expression "conditioned reflex" is intended to express the idea that action is taken not under the influence of current economic policy, but *it is the permanent social circumstances that condition these reflexes*. We do not suggest a fatalistic standpoint. The planner or economic policy-maker can be influenced by his conscious recognition of the fact that as κ approaches 1, so certain marginal social costs grow. If he recognizes this – and his recognition may be assisted both by theoretical economics and by comprehensive social statistics – he might become more cautious and moderate in his decisions. And if he becomes so, he will have to "swim against the current". He will have to anticipate that many other planners participating in forming decisions will continue to see and act in line with the four factors listed above.

[14]In section 12.1 we introduced as assumption (1) the modelling "trick" that the benefit and cost functions would represent an intertemporal "average moment". Thus in comparative static analysis the degree of utilization as cause and the cost as effect appear simultaneously, without any time-label. Yet reality does not consist of intertemporal "average moments". A number of today's causes may have their consequences only tomorrow or even later.

12.12. Intersystem comparisons

It has already been stressed that the planner's views, "conditioned reflexes" and behavioral regularities are not determined by his subjective temperament but by the objective circumstances and social conditions affecting him, and by his role in the system. To that we must add that it is, of course, not only the planners' "conditioned reflexes" that drive the economy towards the full utilization of social capacity. All the control mechanisms, causal relationships, interactions, and "vicious circles" examined in detail in Chapters 2–11 exert their effects in the same direction. Such effects arise from the quantity drive and the associated hoarding tendency, the expansion drive, and the insatiable investment hunger accompanying it and, as a consequence of all this, the almost-insatiable demand by firms for inputs, that is chronic suction. Extremely strong tendencies drive the system close to the state $\kappa = 1$ and, though countertendencies exist, they are not too strong.

We shall say that the social capacity utilization in a system is *taut* if it is not far from full utilization, in the sense that steeply rising marginal social costs are already being experienced. Using this terminology we can make the following proposition.

In the socialist system, with the traditional forms of economic management, the social capacity utilization is taut.

In making this claim we have merely reformulated, in the *macro*-terms of the present chapter, the *microeconomic* proposition of Chapter 2, namely that the traditional socialist economy is resource-constrained.

The phenomenon described in the present chapter of steeply rising marginal social costs at a taut utilization of social capacity, and the phenomena presented in Chapters 2–11 of high shortage intensity, frequent and violent hitting of resource and supply-constraints – are all aspects of the *same* group of phenomena. Each is closely related to the virtual absence of mobilizable slack in the system, since it is continually "siphoned off" by certain social processes. The system always arrives – in its submicro-level elementary events – at or close to the limit of its own physical capacity.

Attention must be drawn to the fact that our proposition can only be correctly interpreted if we bear in mind the earlier ideas in the book, as well as assumptions (1)–(7) of the present model. That is to say, what we call "social capacity utilization" is not identical to the weighted average of the instantaneous degrees of utilization of individual resources. Therefore, the proposition just stated does not imply that the socialist economy uses

100 percent of every resource, or even almost 100 percent. There is a significant amount of slack, but it is nonmobilizable. The hitting of physical resource constraints appears at each instant as thousands of submicro-level bottlenecks. These are necessarily accompanied by slack in the complementary resources not usable for anything else at that moment. On average over time many kinds of resource are unused to a considerable extent, and, *at the same time*, the value of κ approaches 1; these are two phenomena which are logically as well as empirically compatible with each other.

This will become even clearer if it is compared with the capitalist economy in its classical state. There κ is much smaller than 1, since for each production activity there is mobilizable slack of all complementary inputs. Increases in output are not restricted by production hitting resource-constraints. The system's inner laws of movement do not permit expansion to go as far as the physical constraints. That is because the system meets effective demand constraints first. Consequently we can state the following proposition.

In the classical capitalist system the social capacity utilization is not taut.

This proposition, as previously in connection with the socialist economy, just reformulates the corresponding proposition from Chapter 2. We have now described in the *macro*-terms of the present chapter the situation described earlier in *micro*-terms in the following way: the classical capitalist system is demand-constrained.

It is not as a precise diagram but rather to outline the problem that we provide fig. 12.6. We show there the normal degree of utilization characteristic of the classical capitalist and traditional socialist economies. The loops to the right and left of these points represent cyclical or irregular fluctuations around the normal degree. Of all the social benefits and costs only *increasing* marginal *costs* are shown in the figure; the vector of these is represented by a single component. The capitalist system reaches the region where some of the marginal social costs start to increase at the peak of its upswing.[15]

As opposed to this, the average degree of social capacity utilization for the traditional socialist economy is already close to full capacity and at the

[15]Fig. 12.6 presents classical capitalism *prior to* the extensive application of Keynesian economic policy. Keynesian economic policy pursued in the long run shifts the normal degree of utilization of the system along the horizontal axis κ, to the right.

The acceleration of inflation is not independent of this change. Let us consider fig. 12.2: as a consequence of a sustained increase in capacity utilization marginal social costs grow more and more steeply for production inputs accounted for in money terms within the firm. This adds to cost-push inflationary forces.

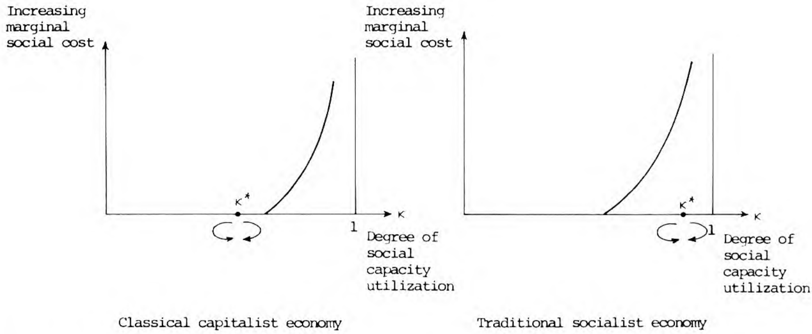


Figure 12.6. Normal degree of social capacity utilization in capitalist and socialist economies.

peaks of investment cycles capacity constraints are met. It is always, even at the lowest points of cycles or irregular fluctuations, deep inside the region of increasing marginal costs.

When we assert that in the traditional socialist economy social capacity utilization is higher than in classical capitalism, we have not yet said anything about the "efficiency" of the two systems. We do not even attempt to provide a summary assessment, since we have just explained that neither social benefits nor social costs can be aggregated. What we can do instead (and this is what this book tries to do) is the following: we want to examine the various consequences of a higher or lower utilization of social capacity in terms of social benefits and advantages, as well as social costs, burdens and disadvantages.

Finally, one more remark. In the above discussion classical capitalism was contrasted with the traditional forms of socialist economy. The question as to how *economic management reforms and changes in growth policy* affect the phenomena examined in the present chapter is left open.

It was already mentioned in section 9.8 that in the past 10–15 years Hungarian central economic policy has generally been anticipating more moderate growth rates than it had been earlier, in the period of overtaut plans. It is increasingly recognized by the planners that forcing expansion has many drawbacks. The "thin lines" in fig. 12.5 became "thicker" in their eyes. Experiences need to be analyzed thoroughly in order for us to infer how substantial this change is, and how permanent; and whether a shift in the normal utilization of social capacity has taken place.

12.13. More on the assumptions

Let us now revert to the assumptions introduced at the beginning of this chapter. We believe that all the general propositions made in the course of this chapter can also be maintained if we relax the simplifying assumptions. It is left to the reader to check this in detail; we shall make but a few remarks.

We have nothing to add to assumptions (1), (2), and (3). It would obviously be desirable to construct a dynamic model instead of relying on comparative static analysis and, instead of considering the system as given – with its physical resources, political-institutional structure, and behavioral regularities – to examine the transformation of systems. The difficulties of the task are quite beyond our powers.

In assumption (4) we have disregarded foreign trade and the use of foreign resources. If this simplifying assumption is removed, physical constraints on the system from within the country do not set an absolute upper limit to expansion. The system may expand beyond domestic capacity. In that case the upper limit depends on the aid- and/or credit-granting propensities of potential donor and creditor countries, as well as on how far the recipient country can accept all the economic and political consequences of aids and/or credits. This was touched upon already in sections 9.10 and 10.6 and will be treated further in Chapter 21, but we draw attention to it at this point, too.

According to assumption (5) the output composition of production is fixed. Obviously, this is a strong abstraction even for analysis in terms of comparative statics. Engel-curves for consumption usually imply that the composition of consumption changes with its total volume. As regards production, all that was said about the nonlinearity of cost curves also implies that output composition and input proportions change as a function of κ.

Although all this could be demonstrated theoretically with a more complicated model, it would add unnecessarily to the difficulty of explaining our point. It was exclusively for expository purposes that – instead of the vectoral measurement usual in the book – we wanted to measure the volume of social output by a single scalar variable. In order to do this, and *without* aggregation using prices, we introduced the assumption of a fixed output composition as an analytical "trick". The consistency of our train of thought required that we should not aggregate here using prices. On the one hand because the present chapter is where we reach the final conclusions of Part I, which only examines nonprice adjustment. And on the other hand because one of the main points of this chapter is precisely that a satisfactory solution of the aggregation problem using prices, and covering all social benefits and costs, is impossible. We do not wish to give the impression that the difficulties of aggregation could be *overcome* by this "trick". The summation of all benefits and costs into one scalar is *impossible* – it is impossible with prices, but just as impossible without prices.

Otherwise, we may guess that taking into consideration the change in output combination would not weaken but strengthen our main conclusions. It has become clear from earlier chapters of the book that the more frequent and troublesome the hitting of resource-constraints becomes (i.e. the nearer the system gets to full utilization of social capacity), the more frequent will be forced substitution and the improvised modification of output combinations, with all its costly consequences. These phenomena were already included among the components of increasing marginal social costs in earlier sections of the chapter (in tacit contradiction to assumption (5)).

According to assumption (6) investment ratios are constant for every value of κ. Historical experience, however, seems to suggest that taut utilization of social capacity is usually coupled with relatively high investment rates.

Besides, cyclical fluctuations are accompanied not only by fluctuations of the volume of investment but also of the investment rate.

Even if this interrelation were built into our model, it would not change our general propositions about the relationships of social benefits, costs, and capacity utilization.

Finally, we add the following to assumption (7), according to which in the framework of comparative static analysis we move along the same iso-friction curve throughout: this assumption must not be lifted. It is particularly important to bear this in mind when comparing systems.

The director of a Hungarian watch factory is justified in saying that he has reached a capacity limit if in the given Hungarian material supply situation and with the Hungarian "normal friction" he is unable to produce more with the machines and personnel of the factory. It would be logically indefensible to say in such a case that surely, the director of a Swiss watch factory would be able to produce more with this factory. It is probably true, but only with the Swiss management practice and organization, the Swiss material supply, and with the Swiss "normal friction". In the terminology of this book, the Swiss clock factory is situated on another iso-friction curve, probably closer to the origin, therefore the "physical resource-constraint" is in a different position.

This line of thought can help in scientific comparisons between different systems. Data sometimes show, at first sight, a paradoxical picture. In country A the capital/output ratio is higher, the ratio of stocks in comparison with current production is higher, there are more hours lost in the daily course of work than in country B. How is this compatible with our claim that it is country A that is a resource-constrained system, while country B is not? It is compatible, because country A is physically unable to produce more *given its own system-specific normal friction*. Here the social capacity utilization is taut, and κ is near 1. At the same time, country B would be physically able to produce more with its own system-specific normal friction. There the social capacity utilization is not taut; κ is well below 1, but it is held down by the low level of effective demand.

SERIES: CONTRIBUTIONS TO ECONOMIC ANALYSIS
EDITORS: D. W. JORGENSON, J. WAELBROECK
HONORARY EDITOR: J. TINBERGEN

A selection of titles

D. J. AIGNER and A. S. GOLDBERGER (Eds.)	*Latent Variables in Socio-Economic Models*
A. AYKAC and C. BRUMAT (Eds.)	*New Developments in the Applications of Bayesian Methods*
B. BALASSA et al. (Eds.)	*European Economic Integration*
R. J. BALL (Ed.)	*The International Linkage of National Economic Models*
H. BARKAI	*Growth Patterns of the Kibbutz Economy*
J. R. BEHRMAN	*Macroeconomic Policy in a Developing Country*
J. R. BEHRMAN, Z. HRUBEC, P. TAUBMAN and T. J. WALES	*Socioeconomic Success: A Study of the Effects of Genetic Endowments, Family Environment, and Schooling*
M. BROWN et al. (Eds.)	*Essays in Modern Capital Theory*
R. H. DAY and A. CIGNO	*Modelling Economic Change: The Recursive Programming Approach*
P. B. DIXON	*The Theory of Joint Maximization*
S. E. FIENBERG and A. ZELLNER (Eds.)	*Studies in Bayesian Econometrics and Statistics*
M. FUSS and D. McFADDEN (Eds.)	*Production Economics: A Dual Approach to Theory and Applications*
J. M. HEINEKE (Ed.)	*Economic Models of Criminal Behavior*
R. J. HERRING and R. C. MARSTON	*National Monetary Policies and International Financial Markets*
B. G. HICKMAN and R. M. COEN	*An Annual Growth Model of the U.S. Economy*
M. D. INTRILIGATOR (Ed.)	*Frontiers of Quantitative Economics, Vol. I*
M. D. INTRILIGATOR and D. A. KENDRICK (Eds.)	*Frontiers of Quantitative Economics, Vol. II*
M. D. INTRILIGATOR (Eds.)	*Frontiers of Quantitative Economics, Vol. III*
L. JOHANSEN	*Production Functions. An Integration of Micro and Macro, Short Run and Long Run Aspects*
G. G. JUDGE and T. TAKAYAMA (Eds.)	*Studies in Economic Planning over Space and Time*
M. C. KEMP and N. VAN LONG (Eds.)	*Exhaustible Resources, Optimality, and Trade*
J.-J. LAMBIN	*Advertising, Competition and Market Conduct in Oligopoly Over Time*
T. C. LEE, G. G. JUDGE and A. ZELLNER	*Estimating the Parameters of the Markov Probability Model from Aggregate Time Series Data*
J. LESOURNE	*A Theory of the Individual for Economic Analysis, Vol. I*
H. LINNEMANN et al.	*MOIRA: Model of International Relations in Agriculture*
S. A. LIPPMAN and J. J. McCALL (Eds.)	*Studies in the Economics of Search*
W. D. NORDHAUS (Ed.)	*International Studies of the Demand for Energy*
P. K. PATTANAIK	*Strategy and Group Choice*
J. S. PETTENGILL	*Labor Unions and the Inequality of Earned Income*
J. D. PITCHFORD and S. J. TURNOVSKY (Eds.)	*Applications of Control Theory to Economic Analysis*
J. M. M. RITZEN	*Education, Economic Growth and Income Distributio*
K. SATO	*Production Functions and Aggregation*
M. SATTINGER	*Capital and the Distribution of Labor Earnings*
J. A. SAWYER (Ed.)	*Modelling the International Transmission Mechanism*
C. R. SCHERER	*Estimating Electric Power System Marginal Costs*
P. TAUBMAN (Ed.)	*Kinometrics: Determinants of Socioeconomic Success Within and Between Families*
W. TRZECIAKOWSKI	*Indirect Management in a Centrally Planned Economy*
J. TSUKUI and Y. MURAKAMI	*Turnpike Optimality in Input-Output Systems*
J. WAELBROECK (Ed.)	*The Models of Project LINK*

1000 AE AMSTERDAM
P.O. BOX 211
NORTH HOLLAND
PUBLISHING CO.
NETHERLANDS

Vol. A ISBN 0 444 85426