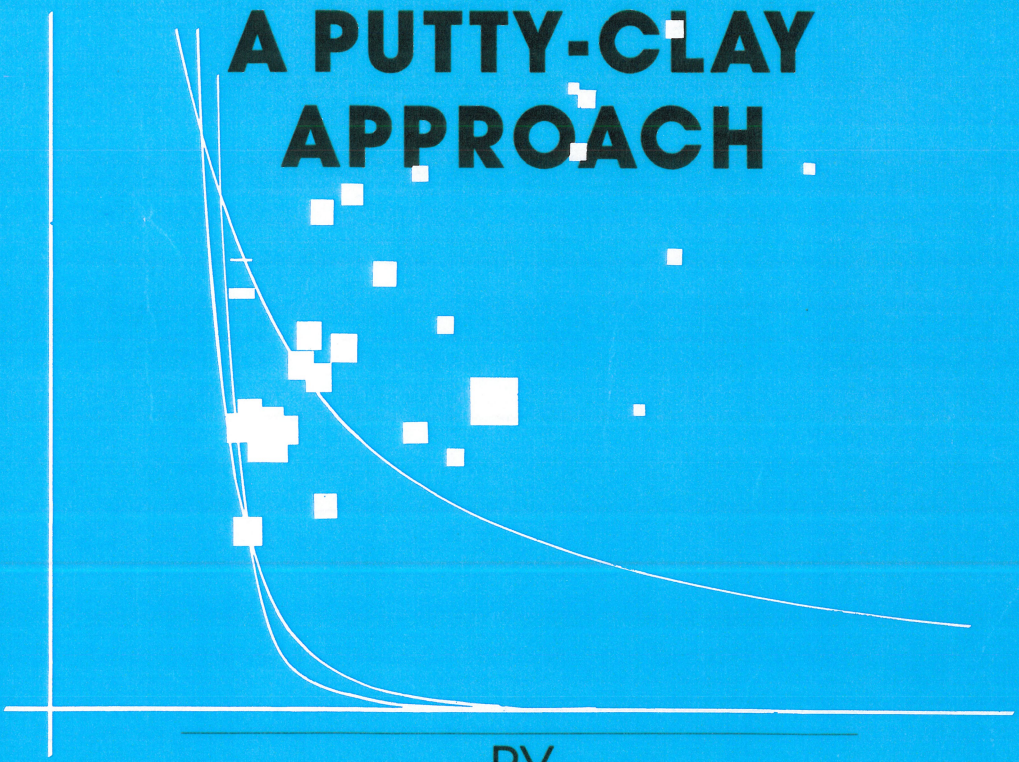
ANALYSES OF INDUSTRIAL STRUCTURE

A PUTTY-CLAY APPROACH



BY
FINN R. FØRSUND
AND
LENNART HJALMARSSON

The Industrial Institute for Economic and Social Research

The industrial Institute for Economic and Social Research

is an independent non-profit research institution, founded in 1939 by the Swedish Employers' Confederation and the Federation of Swedish Industries.

Objectives

To carry out research into economic and social conditions of importance for industrial development in Sweden.

Activities

The greater part of the Institute's work is devoted to long-term problems, especialy to long-term changes in the structure of the Swedish economy particulary within manufacturing industry.

Board

Address

Industriens Utredningsinstitut
Grevgatan 34, 5 tr, S-114 53 Stockholm, Sweden
Tel. 08-783 80 00

Analyses of Industrial Structure: A Putty-Clay Approach

The Industrial Institute for Economic and Social Research

Analyses of Industrial Structure: A Putty-Clay Approach

by

Finn R. Førsund

and

Lennart Hjalmarsson

Distributed by
Almqvist & Wiksell International, Stockholm

Distribution:
Almqvist & Wiksell International, Stockholm, Sweden.

ISBN 91-7204-285-0

Cover picture: Stefan Lehtilä Tecknare AB
gotab Stockholm 1987

Foreword

The study of structural development within different industries has been on the research agenda of the Industrial Institute for Economic and Social Research (IUI), Stockholm, for many years. Within the economics profession a wide variety of approaches to industrial structure analysis has been attempted over the years. However, the lack of a theoretical basis for analysing industrial structure has forced the use of overly simple empirical methods. Førsund and Hjalmarsson offer a unified analytical approach based on the dynamic theory of production.

This study was initiated by IUI many years ago. During its fairly long gestation period several partial presentations have been published in various journals. The time has now come to collect and unify published and unpublished papers into a single volume. We hope that this study will provide a useful reference for both economists in applied economics and researchers in the field of industrial economics.

Stockholm, June 1987

Gunnar Eliasson

Preface

Analyses of industrial structure and productivity growth have a long history in economics. The methods applied, however, have on the whole been fairly simple. One important reason for this has been the lack of a well-grounded, theoretical foundation which has been shown to be empirically relevant to the dynamic study of industrial structure and productive efficiency. Such a theory would have to reflect the fact that an industry usually consists of single production units with different technology and, in the short run, a rigid capital structure. Salter in 1960 provided an important step toward the development of a foundation when he introduced the distinction between best-practice and average productivity. A complete production theory was later worked out by Leif Johansen in the late 1960's. A cornerstone of Johansen's theory is the short-run industry production function. Contemporary with Johansen's work, Aigner and Chu introduced the notion of frontier production, which serves as a suitable tool for the analysis of productive efficiency. Farrell had given a precise meaning and several measurements of productive efficiency in a seminal article of 1957.

This book is an attempt to show how fruitful a production function approach is to the analysis of industrial structure and technical change. Inspired by Leif Johansen, we began our research in the early 1970's. The present study is an outgrowth of many years of work. We have collected and further elaborated upon material which has been published in the following journals: Econometric Reviews, Econometrica, Economic Journal, Empirical Economics, European Economic Review, International Economic Review, Journal of Econometrics and the Scandinavian Journal of Economics (formerly the Swedish Journal of Economics). We thank these journals for permission to use these materials in the present study.

Both of us had the great advantage of Leif Johansen's good advice, encouragement and strong support until his death in 1982. Our debt to him is particularly deep.

Quite a few of our colleagues and students have patiently read and commented upon numerous drafts of this study. They are too many to be mentioned individually. The same holds true for the various secretaries who patiently typed our manuscripts. To all of these people, none of whom is to blame for any remaining errors, we wish to express our gratitude.

A study of this scale would have been impossible without generous funding from different sources. From the outset the project has been supported by the Institute for Industrial and Social Research in Stockholm and the Gothenburg Economic School Foundation, the latter particularly helpful with respect to data and secreterial assistance. At an early stage financial support was obtained from the Swedish Council for Social Science Research and the Norwegian Academy of Sciences. In recent years the project has been supported by the Jan Wallander's Research Foundation, Svenska Handelsbanken, and the Nordic Economic Research Council. We wish to express our sincere thanks for this invaluable support.

Finally, this book has been set using Donald Knuth's computer typesetting system TEX. Professor Knuth's contribution to mathematical typesetting has allowed us the freedom to carry out numerous rounds of revisions and proofreading without sacrificing the final quality of print. For this achievement we are grateful.

Table of Contents

1

Introduction

1.1 The purpose of this study

In recent years we have seen an increased interest in problems of industrial transformation and the structural development of different industries. This is in contrast to the preceding years when a more aggregated view of the process of economic growth was the main topic of concern. The most important reasons for this changing interest are probably empirical, namely, recent dramatic fluctuations in factor prices and the slowdown in productivity growth and rising unemployment observed in many countries. New analytic techniques in economics, however, have also furthered interest in industrial transformation. To some degree the change in emphasis might also be considered as an aftermath of the capital controversy that dominated economics for so many years.

The demand for more detailed analyses of industrial structure seems to be especially great in small open economies. There are several reasons for this. First, these economies in general are characterised by small national markets, insufficient to support even a single plant of minimum optimal scale in many industries. Secondly, most of their industries are exposed to international competition and thirdly, these economies are highly vulnerable to exogenous price shocks on traded goods.

A prerequisite for high productivity growth in small open economies is the ability to rapidly adjust to changing international market conditions. At least in principle, the purpose of an industrial policy is to enhance and promote this market demanded process of structural change. Obviously, the establishing of an industrial policy increases the urgency for a more exact knowledge of the industrial structure and a deeper understanding of the process of structural change in individual industries.

Today, the process of national planning based on large scale econometric models poses questions on how to utilise data from the micro level

of the economy. Leif Johansen in particular stressed

> that much of the modelling of the supply side will fail to come to grips with important problems because it relies too much on smooth, neo-classical formulations of production functions and derived concepts.[1]

To increase the realism and explanatory power of econometric planning models Johansen urged the adoption of the putty-clay approach. However, comprehensive data is required from the micro level of the economy if one is to obtain more reliable econometric results with the implementation of putty-clay concepts.[2]

The purpose of this study is to develop a framework for the analysis of industrial structures. We hope that it will contribute to a better understanding of old but rather vague concepts, e.g., optimal structure, productive efficiency and economies of scale as well as of empirically observed phenomena, such as the efficiency distribution of plants and size distribution of plants. The latter have been difficult to analyse on the basis of the traditional neoclassical theory of production. We believe that our approach as a basis for formulating industrial policy is an improvement over those simple empirical surveys of industrial structure that concentrate exclusively on labour productivity and the size distribution of plants. Further improvement of national economic planning models will require that some representation of the structure of each production sector be introduced into the models.[3]

This study is organised in the following manner. The first part introduces a theoretical framework for a dynamic theory of production based on the putty-clay model, with special emphasis on the concept of optimal structure. Next follows a methodological framework for the empirical estimation of those theoretical production function concepts relevant to the analysis of industrial structure, i.e., the ex ante production function or frontier production function and the short-run industry production function. With respect to the latter, an operational method for the analysis of discrete capacity distributions is developed. In the second part we present some empirical applications of both types of production functions, with special emphasis on the use of short-run industry functions in the analyses of long-run structural and technical change.

---

[1] Johansen [1972], p. 25.

[2] *ibid.* 26.

[3] For an outline of such an approach, see Førsund and Jansen [1983c, 1985].

1.2 A brief historical note

The empirical foundation for the neoclassical theory of production was originally based upon microstudies of production in agriculture. Somewhat paradoxically, the main hypotheses of the neoclassical theory of the firm have had their most successful application at the aggregate level, that is in guiding the allocation of resources in macro models.[4] This situation stems from the fact that in the neoclassical theory it is the outcome of the operation of a perfectly competitive market system and not the actions of the firms per se that is important. The neoclassical production function represents a hypothetical institution operating as a single decision-making unit; it is usually called a firm or an industry. Internal problems of organisation, the decision-making process, the capital structure of micro units, etc., are not within the domain of the theory.

The main objective of neoclassical theory is to predict changes in the supply of outputs and the demand for inputs when the only external variables to which decision-making units react are changing market prices. The neoclassical theory, therefore, is not a suitable tool for analysing problems such as the process of structural change within an industry when firms or plants differ in size and structure with respect to their use of input coefficients or when plants become obsolete. Nor can the neoclassical theory explain what might happen when market size increases, a nonproportional factor price change takes place or embodied technical progress occurs. Structure is only an interesting concept when there is a certain stability, inertia or clayishness in the capital structure of an industry. Without immobility or non-malleability of fixed factors, no structural problem arises.

It is interesting to note that Marx, Schumpeter and Marshall all had interesting comments concerning vintage aspects of industrial structure. Especially Marx, in his *Capital* showed a great interest in the structural development of different industries. His remarks were based upon a genuine awareness of actual empirical investigations of the development of various industries with respect to size, structure, labour productivity and technical progress. Compare the following passages concerning a main theme in a vintage production theory, namely, the existence of different vintages of capital at any one point in time and the gradual transformation of the structure over time:

> The instruments of labour are largely modified all the time by the progress of industry. Hence they are not replaced in their

[4] See, for instance, Johansen [1960].

> original, but in their modified form. On the one hand the mass of the fixed capital invested in a certain bodily form and endowed in that form with a certain average life constitutes one reason for the only gradual pace of the introduction of new machinery etc, and therefore an obstacle to the rapid general introduction of improved instruments of labour. On the other hand competition compels the replacement of the old instruments of labour by new ones before the expiration of their natural life, especially when decisive changes occur.[5]

And on obsolescence:

> But in addition to the material wear and tear, a machine also undergoes what we may call a moral depreciation. It loses exchange value, either by machines of the same sort being produced cheaper than it, or by better machines entering into competition with it.[6]

According to Schumpeter the essence of capitalism is its creative destruction, by which he means the evolutionary process in capitalism driven by innovations. This process of innovation incessantly revolutionises the economic structure *from within*, incessantly destroying the old structure, incessantly creating a new one.[7] The Schumpeterian process implies that any industrial structure which happens to exist at a particular moment will rapidly become obsolete and consequently be abandoned.[8]

In his three volumes of *Capital*, Marx's theory of production stands out as a typical vintage theory.[9] However, the term *quasi-rent* comes from Marshall's *Principles of Economics:*

> When any particular thing, as a house, a piano, or a sewing machine is lent out, the payment for it is often called *Rent.* And economists may follow this practice without inconvenience when they are regarding the income from the point of view of the individual trader. But, as will be argued presently, the balance of advantage seems to lie in favour of reserving the term Rent for the income derived from the free gifts of nature, whenever the discussion of business affairs passes from the point of view of the

[5] *Capital*, 11:8, p. 174.

[6] *Capital*, 1:15, p. 381.

[7] See Schumpeter [1950], p. 83.

[8] See Elliott [1980].

[9] This is further discussed in Hjalmarsson [1975]. See also Elliott [1980].

> individual to that of society at large. And for that reason, the term *Quasi-rent* will be used in the present volume for the income derived from machines and other appliances for production made by man.[10]

Compare the following section on obsolescence:

> It is of course just as essential in the long run that the price obtained should cover general or supplementary costs as that it should cover prime costs. An industry will be driven out of existence in the long run as certainly by failing to return even a moderate interest on capital invested in steam engines, as by failing to replace the price of the coal or the raw material used up from day to day. ... So an industry may, and often does, keep tolerably active during a whole year or even more, in which very little is earned beyond prime costs, and the fixed plant has "to work for nothing". But when the price falls so low that it does not pay for the out of pocket expenses during the year for wages and raw material, for coal and for lighting, etc., then the production is likely to come to a sharp stop.
>
> This is the fundamental difference between those incomes yielded by agents of production which are to be regarded as rents or quasi-rents and those which (after allowing for the replacement of wear-and-tear and other destruction) maybe regarded as interest (or profits) on current investments.[11]

In his theory of production, however, Marshall does not develop these vintage aspects further. Instead his analysis is based on the idea of the *representative firm.*

Still other, more disparate, comments or traces of vintage notions can be found in the literature. In Mitchell [1937] the term "best current practice" is found in a discussion about what would be the potential increase in output if all existing equipment could be transformed to best-practice equipment:

> In 1933, twenty-eight engineers of experience in various industries were persuaded to submit estimates of how much the aggregate output of all industries might be increased simultaneously with existing equipment and methods, provided a ready market could

[10] Book II, Ch. IV pp. 62–63.

[11] Book V, Ch. IX, p. 349.

> be assured for the products. More than half of the estimates ran above 25 per cent. Asked what increase might be expected if the equipment and management of all industries were "brought to the level of the best current practice" half of the engineers have estimates of 60 per cent or more. And if the engineers are right, these increases might be doubled or trebled by bringing equipment and management in all enterprises abreast of the best current practice. (Mitchell [1937], p. 119.)

In the history of Scandinavian economics there is a long tradition of interest in problems of industrial structure. In fact, as early as 1918 the Swedish economist Eli Heckscher in a book on Swedish industrial problems, introduced a diagram in which the firms' current average costs were sorted in increasing order. On the basis of such a diagram, Heckscher carried out an analysis of the impact on industrial structure of tariff changes.[12] Other Swedish economists, for example, Åkerman and Svennilson, should also be mentioned. In a study from 1931, Åkerman investigated the difference between the best-practice and average productivity of labour for Swedish saw mills. He showed that during the period 1923–26, the input coefficients of labour for the most modern plants were only 50 per cent of that for the average of the industry. The distance between best-practice and average practice was also discussed in an article by Svennilson [1944]. In the light of our own study, Svennilson's article is most interesting. It includes a thorough analysis of the determinants of the rate of growth of industrial productivity and a simple model from which "ratios of inoptimality" are calculated. These ratios of nonoptimality show the percentage ratio between the average and best-practice input coefficients for labour as a function of the rate of growth of production, the physical lifetime of equipment and the input-coefficients of labour for each vintage of capital. A main point is the relationship between the rate of growth of production and the rate of productivity growth, which is also treated in an empirical analysis of Swedish industries.

With respect to the present state of the dynamic theory of production one has to distinguish between at least two different approaches:

1. One approach is a microdynamic theory of production focusing on growth and investment decisions of the firm. This approach originates

[12] We are grateful to Leif Johansen for drawing our attention to Heckscher's book. The diagrams of sorted average costs will be called Heckscher diagrams. Similar distributions of factor input coefficients which appeared in Salter [1960], are usually referred to as Salter diagrams.

from Marshall and the neoclassical theory of production. Here the behaviour of a representative firm is studied with various types of cost of adjustment models or other steady-state growth models of the firm.[13]

2. The second approach is the dynamic theory of production based on assumptions about ex ante versus ex post substitutability and embodied technological progress. Here the interest is not so much in the average firm, but rather in the whole structure of the industry as regards input coefficients and the size distributions of the micro units. In our study we are solely concerned with this approach. More explicitly, we base our analysis of industrial structure on the production function concepts introduced in Johansen [1972].

Of course, there are close links between these two directions as represented by, for instance, studies of the investment decisions of firms based on capacity expansion models under putty-clay conditions.[14] Primarily, we are not interested in the determinants of investment decisions of firms, but rather in the consequence for industrial structure and structural change of those decisions within a putty-clay framework.

1.3 The concept of structure

The industrial policy in Scandinavia has put a great emphasis on the productive efficiency of the business sector, and on the so-called "structural rationalisation" of various slow productivity growth industries with eroded competitiveness. Thus, structural rationalisation policy has been directed towards a more efficient utilisation of resources, such as labour, by squeezing out the less efficient firms.

The theoretical underpinnings of this policy is not directly related to the traditional theory of industrial organisation with its emphasis on allocative efficiency and anti-trust policy. Instead it is more closely related to dynamic production theory.

The concept of structure has many different meanings in economics. Generally speaking, structure refers to the distribution of some typical

[13] See, e.g., Gould [1968], Lucas [1967] and Nickell [1978]. For a survey, see Söderström [1976].

[14] See, e.g, Nickell [1978] and Freidenfelds [1981].

characteristics of the industry such as distribution of wages, profits, factor productivities, size, market shares, R & D expenditures, advertising expenditures, assets and age of equipment.

Structure and structural change may pertain to different levels of an economy, from the individual firm, or parts of it, to the economy as a whole. We shall be concerned mainly with the industry level. To be a useful concept here, structure must be related to some degree of inertia, and changes in the structure should not be without costs. It is almost meaningless to talk about structure in this sense in a neoclassical world of smooth substitution possibilities and choices of capacity.

Usually we think of structure when there is inertia in the capital structure of an industry. Existing equipment and buildings cannot change their productive capacity without costs. In the case of embodied technological progress and changing input prices both time and investments are necessary to transform a capital stock.

An analysis of industrial structure requires a dynamic theory of production. Various models can be formulated, depending on the degree of inertia in the capital structure. One of the most important models generating stability and inertia in the capital structure is the putty-clay model, which is further discussed in the next section. Accordingly, the elements of structure to which we will pay particular attention are the distribution of input coefficients (input per unit of output) and the output capacity for the micro units of the industry. Depending on the purpose, the micro units can be firms, plants or individual pieces of equipment.

Textbooks in industrial organisation are dominated by the *structure-conduct-performance model*, originating from Mason, Bain and others. There are four main features of this model:

(i) The characteristics of market structure are considered as exogenous, and market conduct and market performance are explained by the market structure.
(ii) The term structure is defined by many structural variables with reference to a single seller or buyer.
(iii) A main concern in the analysis is the strategic market behaviour of the single agent, or group of agents, who are similar in some respects.
(iv) The analysis is based on the theory of market behaviour and market power under various market forms.

Thus, in our study the meaning of structure is different from that of the traditional industrial organisation theory, in which structure usually refers to

market structure and where the main elements are concentration, product differentiation and barriers to entry.

1.4 A dynamic theory of production

The putty-clay growth model

Concurrent with the rather rapid and stable economic growth experienced by most advanced industrial countries in the fifties and sixties was a rapid development of theories of economic growth. One of these theories was the putty-clay model initiated by Johansen [1959]. In this model equipment with differing factor ratios can be designed, but once the equipment is constructed the factor ratio remains constant. Soon after the publication of Johansen's article, Solow [1962a, 1962b, 1963], Phelps [1963], Kurz [1963], Kemp and Thanh [1966] and Bliss [1968] all contributed to the extension and perfection of this type of model. These models, characterised by ex ante factor substitutability and ex post nonsubstitutability, were aptly called "putty-clay" by Phelps.[15]

In the putty-clay model there are as many different kinds of capital goods as there are points of time. These different capital goods are called vintages. A unit of a capital good of a given vintage will provide a certain capacity for producing output, and it will require a fixed unit of current inputs per unit of output (input coefficients). These characteristics remain unchanged throughout the life of the capital good. Technical progress then implies that capacity of a later vintage will always be more efficient than that of an older vintage.

One of the advantages of this model is that it brings obsolescence of capital into the analysis. Even if capital goods last forever so far as their physical characteristics are concerned, they will become economically useless, not because they wear out but because they become incapable of covering their costs, i.e., of earning positive rents. This possibility was excluded from the neoclassical model, even when technological progress was present, because all capital goods, old and new alike, received equal shares of technological progress. Since capital was homogeneous in this model, no single capital good could become obsolete unless the entire capital stock became obsolete.

[15] For a pedagogical exposition, see Solow [1970].

With respect to capacity and the productivity of current inputs (input coefficients) the whole structure of capital goods are brought into the picture with the putty-clay model. The structure of capital goods generates a profile of quasi-rents analogous to the rent of land in the Ricardian theory. The quasi-rent for a piece of capital is defined as the economic surplus after deduction for current operating costs. Older (less efficient) capital is at any time earning a lower quasi-rent per unit of capacity because it pays the same prices for current inputs as newer capital, but is less productive with current inputs. When real wages rise over time, the wage bill of an old factory rises and its quasi-rents diminish. Eventually it becomes a marginal no-rent factory. If wages then increase only slightly, the marginal factory goes out of business. It has become obsolete, not because of any reduction in its efficiency, but because rising operating costs have rendered it incapable of covering its own variable costs of production.

The putty-clay model successfully combines microeconomic investment theory on the one hand, and growth theory on the other. In one respect the model presents a highly complicated structure because there is no longer a meaningful aggregate stock of capital whose numerical magnitude can be examined. Nevertheless, the long-run properties of the steady state development are similar to those of the standard neoclassical models. In the steady state, the economic lifetime, T, is constant; each successive vintage of capital becomes obsolete after T years of operation. Outside the steady state, the economic lifetime varies from one vintage to the next.

The putty-clay production theory

The traditional neoclassical theory of production, with its assumption of smooth (costless) possibilities of substitution and choice of optimal scale, is a suitable tool for the analysis of the long-run development of industrial structure at an aggregate level. However, it is not suitable for the analysis of short-run, or medium-term problems of industrial structure within an industry.

In recent years most work on production theory has occurred within the dual approach, pioneered by Shephard [1953]. This approach links cost and production functions. A comprehensive treatment of both theoretical and empirical analyses of duality relationships is found in Fuss and McFadden [1978]. Even though the putty-clay framework is applied to some extent in the latter study, a more direct use of putty-clay, with emphasis on the choice of technology when deciding upon new capacity and concern over the entire process of structural change in an industry, can be

found in an approach originating with Salter [1960]. A cornerstone in this alternative development of production theory was Johansen's *Production Functions* [1972], in which a dynamic theory of production is developed through an integration of micro and macro and of short and long-run aspects. The result is a production theory modelling the development of a whole industry producing a homogeneous output. Empirically, this framework provides the possibility of an empirical insight into the structural change of an industry which is deeper and more relevant than, for example, that obtained by an analysis based on the traditional "estimated average production function".

The crucial assumptions in putty-clay production theory concern the substitution possibilities with regard to factor proportions and capacity (i.e., full substitution possibilities ex ante, but fixed factor proportions and capacity ex post) and the emphasis on embodied technological progress, leading to different vintages of capital and a gradual transformation of the structure over time. The main ideas were already proposed in Johansen [1959], and closely related ones were found in Salter [1960] with his distinction between best-practice and average-practice productivity.

Johansen distinguishes between four different production function concepts:

1. *Ex ante function at the micro level.* This is the production function which exists at the moment of investment and from which the choice of technique is made. We may characterise it as a traditional production function with continuous substitution possibilities.
2. *Ex post function at the micro level.* This is characterised by fixed-production coefficients and is the relevant production function after the moment of investment.
3. A *short-run industry production function* built up from the ex post functions of the micro units.
4. A *long-run industry production function* which is closely connected with the ex ante function.

If we consider an industry consisting of a certain number of micro units, the short-run industry production function is established by maximising output for given levels of current inputs. Thus, it corresponds to the basic definition of a production function when the industry is regarded as one production unit, unlike the traditionally estimated function for an industry. The traditional approach is based on the notion of the representative firm, i.e., when estimating an average industry function it is assumed that all micro units have the same underlying production technology, except for

a random error term. In contrast, the short-run function explicitly recognises that the technologies of the individual micro units differ. It utilises the information about these different technologies to establish by explicit optimisation the relationships between the aggregate industry output and inputs. Thus in a putty-clay world the short-run function is the true function for the industry as a whole. Due to the unique relationship between actual technologies and the short-run function, the latter and its derived relationships provide us with a well-defined concept of industrial structure.

A series of short-run industry production functions over time are connected by way of the ex ante production functions. The ex ante function can be regarded as a choice-of-technique function for the construction of an individual micro unit. The short-run industry production function reflects both the history of ex ante functions over time and the actual choices made from these ex ante functions. Production at any time must be compatible with the short-run function.

The factors studied within the short-run function are limited to current inputs. Fixed factors, such as capital, only determine the capacity of the individual micro units. In the ex ante function all factors are variables. The ex ante function at the micro level is the production function existing at the moment of investment and from which the choice of technique is made. We characterise it as a traditional production function with continuous substitution possibilities. Each production unit has been "extracted" from the ex ante function that existed at the time of construction. The short-run industry production function reflects both the history of ex ante functions over time and the actual choices made from ex ante functions. The ex ante function can be derived from engineering knowledge, or estimated as a frontier production function.[16] In the ex ante case the requirement for information about technical relationships is much greater than for the short-run function.

1.5 The scope of this study

The purpose of this study is to develop a framework for the analysis of industrial structures applicable to several industries. Chapter 2 analyses how specific structures may be generated. Various vintage models are presented and the concepts of optimal structure and optimal structural change are discussed. The chapter also utilises the vintage theory of production

[16] See, e.g., Eide [1979] and Chapter 4, respectively.

to shed some new light on old problems in economics. In addition, the implications of a vintage model for the size distribution of micro units are developed. Monopoly welfare gains, scale efficiency and the costs of decentralisation are analysed with reference to Williamson's trade-off model. Reading this chapter is not necessary for the understanding of the rest of the study.

Chapter 3 analyses the concept of efficiency on the basis of a frontier production function, generalising Farrell's measures of productive efficiency. Technical change is characterised and Salter's measures are generalised to nonhomogeneous production functions. Studies of frontier or of best-practice production functions based on data for micro units, combined with studies of productive efficiency, have to some extent replaced the traditional average production function estimations usually carried out on highly aggregated data. Chapter 4 surveys empirical approaches to the estimation of frontier production functions.

Chapter 5 presents the short-run industry production function. Johansen's approach is developed into an operational framework for discrete capacity distributions including a special algorithm for the computation of the short-run industry production function.

Chapter 6 summarises the main results from the empirical applications. Chapter 7 is an empirical study of the Swedish dairy industry with special emphasis on technical progress and productive efficiency. Chapter 8 studies technical progress and structural change in the Swedish cement industry 1955–1979 on the basis of the development of the short-run industry production function. In Chapter 9 a similar analysis is performed for the Swedish pulp industry 1920–1974, and in Chapter 10 for Swedish pig-iron production 1850–1974. Finally, Chapter 11 deals with the Norwegian aluminum industry on the basis of a short-run industry production function analysis for the period 1966–1978.

2

Optimal Structural Change and Related Problems

2.1 Introduction

This chapter is concerned with the dynamic process of structural change in an industry producing a homogeneous product. We are especially interested in the origins of differences in the structure with respect to capacity and input coefficients. Hence, we consider the development of the structure over time through the process of choosing new techniques and investments from the ex ante production function, and through the closing down of old equipment that has begun to earn negative quasi-rents due to changes in product and factor prices.

In Section 2.2 we look at the investment and scrapping decisions of a single micro-unit. Next, in Section 2.3 we consider the industry as a whole. There the concepts of optimal structure and optimal structural change are defined and illustrated. In Section 2.4 a capacity expansion model for an industry is presented. The model is based on putty-clay assumptions and economies of scale in the ex ante production function. The industrial structure with respect to the choice of technology and size of different plants is then studied. The entire size distribution of plants derived from this capacity expansion model is examined more closely in Section 2.5, where we also refer to older explanations of empirically observed, generally skewed, size distributions. In Section 2.6, we then utilise the same model to throw some new light on another old topic, originally discussed in Williamson [1968], namely, the trade-off between the exploitation of economies of scale and increases in industrial concentration.

2.2 The vintage model

As a first step in the theoretical analysis we will present here a very simple putty-clay production model for a single firm. The model is the vintage analogy to the standard profit-maximising model in the neoclassical theory of the firm. From the model we obtain the basic putty-clay results about technology choice and scrapping criteria.

Basic assumptions

The two aspects of the vintage model to be analysed here are:

1. *The investment decision:* What factor ratios are chosen? What volume of production is planned? What determines the planned lifetime of the plant?
2. *The scrapping decision:* What are the criteria for scrapping?

Production possibilities at the planning stage, are described by the ex ante production function:

$$x(\nu,\nu) = f_\nu\big(v_1(\nu,\nu),\ldots,v_n(\nu,\nu),K(\nu,\nu)\big) \tag{2.1}$$

where the first argument denotes time and the second argument denotes vintage.

In (2.1) we have

$f_\nu(\cdot)$ = ex ante production function at time ν. The production function is assumed to have the "usual" properties such as continuous differentiability and positive, but falling marginal productivities in the substitution region.

$x(\nu,\nu)$ = planned production at time ν with capital of vintage ν.

$v_i(\nu,\nu)$ = planned use of current input i, $(i = 1,\ldots,n)$ at time ν with capital of vintage ν.

$K(\nu,\nu)$ = planned capital investment at time ν.

Two time indices are used in the description of the current ex post production possibilities. Ex post we have a limitational law in the current inputs:

$$\begin{aligned} x(t,\nu) &= \min\left\{\frac{v_1(t,\nu)}{\xi_1(\nu)},\ldots,\frac{v_n(t,\nu)}{\xi_n(\nu)},\bar{x}(t,\nu)\right\} \\ x(t,\nu) &\in [0,\bar{x}(t,\nu)] \end{aligned} \tag{2.2}$$

In (2.2) we have

$x(t,\nu)$ = production at time t with capital of vintage ν.
$v_i(t,\nu)$ = current input i in use at time t together with capital of vintage ν, $i = 1, \ldots, n$.
$\xi_i(\nu)$ = constant input coefficient for input i ($i = 1, \ldots, n$), valid for vintage ν.
$\bar{x}(t,\nu)$ = maximum production capacity for a micro unit at t with capital of vintage ν.

Input coefficients ex post emerge as a "freezing" of planned input coefficients $v_i(\nu,\nu)/x(\nu,\nu)$, where we assume the planned magnitudes refer to full utilisation of capacity. Capacity is determined by the realisation of planned capital, $K(\nu,\nu)$. In order to formulate the economic problems involved we must make certain assumptions with regard to the prices of the applicable variables in the production function. The decision-making unit must form expectations about price developments for the entire period of planned operation. Current inputs might be, for example, labour, raw materials and energy. Fixed inputs, i.e., amounts determined at the time of investment, can for instance be divided into structure, capital equipment and transport equipment. These categories are lumped together in a variable called capital.

In order to simplify our analysis we shall assume that there exists perfect capital markets and no second-hand markets for capital equipment, or that the equipment has no positive scrap value. We shall also ignore maintenance costs. Only the initial capital price has any significance. The price is expressed per physical unit of capital. In this context the unit chosen as capital is irrelevant.

The investment decision

We assume that the investor has certain price expectations, i.e., in making his decisions he does not take into account the fact that probability distributions for prices exist at a given time. The investor makes up his mind what he expects future prices to be and then reckons that these expected values will be realised. In the model expected prices are assumed to be exogenous. We also assume price-taking behaviour, and that the firm maximises the present value of profits over the lifetime of the capital. The economic life-span of capital is endogenous and appears as a variable to be determined. The physical lifetime of capital is always greater than the optimal lifetime, which is determined by a maximisation problem. Finally, we

assume that at a certain point in time an evaluation is made as to whether an investment should be carried out, and if so, how much is to be invested. (In the event of technical improvements, a separate problem is posed by the need to determine the optimal point in time to invest if factor supplies are limited.)

The discounted profit function is written as follows:

$$\pi(0) = \int_{t=0}^{T} e^{-rt}\left(p(t)x(t,0) - \sum_{i=1}^{n} q_i(t)v_i(t,0)\right) dt - q_K(0)K(0,0) \quad (2.3)$$

The discount rate of the decision-making unit is r, whereas T is the yet to be determined lifetime of the plant. The discount rate r can be interpreted as the investor's minimum demand for capital return. The project will be undertaken if the present value of profits is nonnegative. A positive present value requires a capital return greater than r. Expected factor prices at time t are $q_i(t)$, $i = 1, \ldots, n$, K. $p(t)$ is the expected product price at time t. We assume that r is a real discount rate and that prices are interpreted as constant, i.e., that current expected prices are deflated according to the expected development of the price index. The price index is the one the investor finds most relevant, for example, the consumer price index. The time of investment has been set conventionally at 0, ignoring the fact that it actually takes time to carry out an investment plan. The plant is planned so as to be operated at full capacity. We ignore the possibility that the decision-making unit may foresee fluctuations in demand. It follows from the putty-clay assumption that the plant's factor ratios are frozen at the level established during the time of investment. For all points of time t, the following therefore applies:

$$x(t,0) = x(0,0) = x \quad \text{and} \quad v_i(t,0) = v_i(0,0) = v_i$$

The indexes for the current time period and vintage are supressed when no confusion arises. Equilibrium conditions for every input are found by making the partial derivatives of the profit function (2.3) equal to zero. The ex ante production function is inserted in the integrand in (2.3). For the current inputs we obtain:

$$\begin{aligned} \frac{\partial \pi(0)}{\partial v_i} &= \int_0^T e^{-rt}\left[p(t)\frac{\partial f_0}{\partial v_i} - q_i(t)\right] dt \\ &= \frac{\partial f_0}{\partial v_i} p(0,T) - q_i(0,T) = 0 \end{aligned} \quad (2.4)$$

where

$$p(0,T) = \int_0^T e^{-rt} p(t)\, dt, \qquad q_i(0,T) = \int_0^T e^{-rt} q_i(t)\, dt \tag{2.5}$$

expresses the present value of prices. We see that this necessary first-order condition for maximising the present value of profits is analogous to the condition found for the traditional static equilibrium. Instead of referring prices to a point in time, however, we operate with the present value of the expected price series.

In place of present values we could operate with average prices over the planned life-span. These average prices for the price functions $e^{-rt}q_i(t)$, $e^{-rt}p(t)$ are respectively:

$$\bar{q}_i = \frac{\int_0^T e^{-rt} q_i(t)\, dt}{T} \qquad \bar{p} = \frac{\int_0^T e^{-rt} p(t)\, dt}{T} \tag{2.6}$$

Condition (2.4) may then be expressed as

$$\frac{\partial f_0}{\partial v_i} \bar{p} = \bar{q}_i \qquad i = 1, \ldots, n \tag{2.7}$$

The value of the initial marginal productivity (ex ante), calculated at the average product price, must be equal to the average factor price. Note that generally the expected (discounted) prices will be equal to their average values defined in terms of (2.6) at different points in time. If we assume a monotonically rising ratio between factor prices and the product price, the value of the marginal productivity will be equal to the factor price at only one particular point in time; this point will generally be different for the different factors and also different from the point of time at which the average price (2.6) actually occurs. With monotonic price developments of this kind, the value of the ex ante marginal productivities is higher than the factor prices during the first part of the plant's active life, and lower during the later part.

Let us assume that prices change on the basis of fixed percentage rates:

$$q_i(t) = q_i(0)e^{\gamma_i t}, \quad p(t) = p(0)e^{ht} \tag{2.8}$$

The necessary first-order condition can then be written as

$$\frac{\partial \pi(0)}{\partial v_i} = p(0)\frac{\partial f_0}{\partial v_i}\frac{1 - e^{-(r-h)T}}{r-h} - q_i(0)\frac{1 - e^{-(r-\gamma_i)T}}{r-\gamma_i} = 0 \tag{2.9}$$

The first term on the right-hand side is the present value of the marginal productivity of factor i. The rate of discount in this present value calculation is the difference between the discount rate of the firm and the growth rate of the product price. If this difference is zero, the present value of the factor's marginal productivity during the lifetime of the vintage is equal to the value of marginal productivity in the year of commencement, multiplied by the lifetime. The second term on the right-hand side is the present value of the costs of input i. The rate of discount in this calculation is the difference between the firm's discount rate and the growth rate of the factor price. In order to arrive at a comparison between this equilibrium and the traditional static equilibrium, the first-order conditions may be arranged in the following way:

$$p(0)\frac{\partial f_0}{\partial v_i} = q_i(0)\frac{r-h}{1-e^{-(r-h)T}}\frac{1-e^{-(r-\gamma_i)T}}{r-\gamma_i} \tag{2.10}$$

On the left-hand side we now have the value of marginal productivity in the year of commencement. If we choose the year of commencement as a basis for comparison of a time-extensive adjustment, then on the right-hand side we have the factor price in the year of commencement multiplied by two terms. These terms can be interpreted as the inverse value of the discounted product price and the discounted factor price, respectively. We may also consider the product of these terms as the relationship between the average value of the factor prices over a period of time T and the average value of the product price for period T.

Static equilibrium for the year of commencement implies equality between the value of the marginal productivity and the factor price in that year. This means that if the value of the product of the "discount terms" in (2.10) is less than 1, it will prove optimal to make greater use of the factor. The factor price's "correction term" will be less than 1 if the product price's growth rate is greater than the factor price's growth rate. (This will be apparent by inspecting the integrals in (2.5) when (2.8) is inserted.)

In the vintage model there will, in the regular case, only be one point in time at which a factor is rewarded with the value of its marginal productivity. If the product price rises more rapidly than a factor price, the current value of the marginal productivity will be less than the factor price during the initial part of the period, and greater than the factor price during the later part of the period.

If a factor price rises more rapidly than the product price, the correction term in (2.10) will be greater than 1. During the year of commence-

ment less of the factor is used than would have been the case in a static equilibrium. The factor initially produces a higher yield than the current value of marginal productivity, but during the later part of the period it provides a current yield that is lower than the current value of marginal productivity.

For adjustment of capital, differentiation of the present value of profits, (2.3) gives us the following first-order condition:

$$\frac{\partial \pi}{\partial K} = \int_0^T e^{-rt} p(t) \frac{\partial f_0}{\partial K} \, dt - q_K(0) = 0$$

equivalent to

$$\frac{\partial f_0}{\partial K} p(0,T) = q_K(0), \tag{2.11}$$

or

$$\frac{\partial f_0}{\partial K} \cdot \bar{p} \cdot T = q_K(0) \tag{2.12}$$

In order to ensure the right dimension on both sides of (2.12) we must multiply the average value of the capital's marginal productivity by the total utilisation time. Using price forecast (2.6) we get:

$$p(0) \frac{\partial f_0}{\partial K} \frac{1 - e^{-(r-h)T}}{r-h} = q_K(0) \tag{2.13}$$

The present value of capital's marginal productivity is made equal to the price per capital unit at the time of investment. The rate of discount in this present value calculation involving an exponential price forecast is, as for the other factors, the difference between the firm's discount rate and the growth rate of the product price.

It now remains to determine the planned lifetime of the plant. Once again we differentiate the present value of profits (2.3), but now with respect to the lifetime T. This is the upper limit for the integral, so we insert this value in the integrand:

$$\frac{\partial \pi}{\partial T} = e^{-rT} \left(p(T)x - \sum_{i=1}^{n} q_i(T) v_i \right)$$

The first-order condition can be written in the following form:

$$\sum_{i=1}^{n} q_i(T) \frac{v_i}{x} = p(T) \tag{2.14}$$

On the left-hand side we have the input coefficients $v_i/x = \xi_i(0)$ for the current inputs. Given our assumptions, these will be constant for every input over the lifetime of the plant. Condition (2.14) tells us that the planned lifetime T is determined in such a way that the total expenses for current factors per unit produced at time T are equal to the product price at time T. This condition is generally formulated in terms of the *quasi-rent.*[1] When the quasi-rent is zero, the costs for current inputs per unit produced are equal to the product price.

The existence of a solution for the lifetime is dependent on the properties of the expected price developments, i.e., sooner or later the quasi-rent must be permanently non-positive. Empirically it is often observed that the factor prices on the "average" rise more steeply than product price.

It is important to note that Equations (2.4), (2.11) and(2.14)must be solved simultaneously.

Throughout this section it is assumed that we obtain a solution for the plant's lifetime such that the economic lifetime is less than the technical lifetime. It is possible, however, that due to price fluctuations a solution might include several periods with zero levels of activity. If there are no costs involved in allowing units to remain on standby, then they might be kept idle until their quasi-rents are once again positive. The solution for the economic lifetime is then the largest T satisfying Equations (2.4), (2.11) and (2.14), with the technical lifetime as the upper limit. When integrating, years with zero capacity utilisation are eliminated. With the more realistic assumption of positive scrap-value included, the "storing" of production capacity becomes a decision in which expected future earnings are weighed against the loss of interest that could be earned were the scrap-value to be realised. Any direct "storage costs" can easily be included in an analysis of this kind.

In the above we have consistently referred to the expected point in time for permanently closing down. This point in time must be derived *ex ante* in order to determine the period over which to integrate. What will the closing-down criterion be *ex post*? It is implicit under price-taking behaviour that the investor has no difficulty in acquiring inputs. It would then be worthwhile to carry out full production with the respective units so long as actual quasi-rents are nonnegative. Under this assumption it is not relevant to speak of transferring, for example, labour to more modern equipment where productivity is higher, before the quasi-rent is zero. In the case of restrictions on factor supplies we may say that the market prices

[1] See Section 1.2.

include "shadow price mark-ups", such that the the consideration of quasi-rent remains the closing-down criterion, but now at "constraint-adjusted" factor prices.

The cost function

Analogous with the consideration of the present value of profits, the relevant cost function in the putty-clay model comprises initial capital costs plus discounted current costs calculated over the time when the plant is in use:

$$C(0) = \int_0^T e^{-rt} \sum_{i=1}^{n} q_i(t) v_i(t,0)\, dt + q_K(0) K(0,0) \tag{2.15}$$

The present value of costs for a given output level with respect to inputs gives us necessary first-order conditions of the same kind as (2.4) and (2.11), but with the Lagrange parameter instead of $p(0,T)$. This parameter belongs to the production function constraint. Since we assume full utilisation of capacity throughout the period of use, we need only one production function constraint, equation (2.1) at a given output level.

Analogous with static production analysis, the factor quantities minimising the present value of costs can, given the production constraint, be expressed as a function of the given output level, factor price functions $q_i(0,T)$ and the capital price $q_K(0)$ (provided that a regular minimum exists).

The output adjustment rule, price equal to marginal cost, can now be obtained by maximising the present value of total sales minus the present value of costs with respect to output, where we assume that the cost-minimising factor quantities have been inserted in the cost expression $C(0)$, yielding the optimised cost function C_0

$$\max_x \int_0^T e^{-rt} p(t) x\, dt - C_0(x) \tag{2.16}$$

The constant factor price functions are here included in the functional form $C_0(\cdot)$. The necessary first-order condition will be

$$\int_0^T e^{-rt} p(t)\, dt \equiv p(0,T) = \frac{\partial C_0}{\partial x} \tag{2.17}$$

The present value of the product price is then equal to the present value of marginal costs, defined by a change in the initial output level x.

By inserting the profit-maximising factor amounts in the cost expression (2.15), i.e., by using $v_i(t,0) = v_i$, and inserting the factor prices from (2.4) and (2.11), we obtain the following expression for the cost function:

$$\begin{aligned} C_0 &= p(0,T)\left(\sum_{i=1}^{n}\frac{\partial f_0}{\partial v_i}v_i + \frac{\partial f_0}{\partial K}K\right) \\ &= p(0,T)\varepsilon_0 \cdot x \end{aligned} \tag{2.18}$$

By utilising (2.17), we see that the present value of costs can be expressed as the product of present-value marginal costs, the elasticity of scale in the ex ante function ε_0, and the output level. If we make use of (2.18) the investment decision can be expressed in terms of the following condition:

$$\sum_{i=1}^{n}\frac{v_i}{x}q_i(0,T) + q_K(0)\frac{K}{x} = p(0,T)\varepsilon_0 \tag{2.19}$$

or expressed by means of average prices:

$$\sum_{i=1}^{n}\frac{v_i}{x}\bar{q}_i + \frac{q_K(0)}{T}\frac{K}{x} = \bar{p}\cdot\varepsilon_0 \tag{2.20}$$

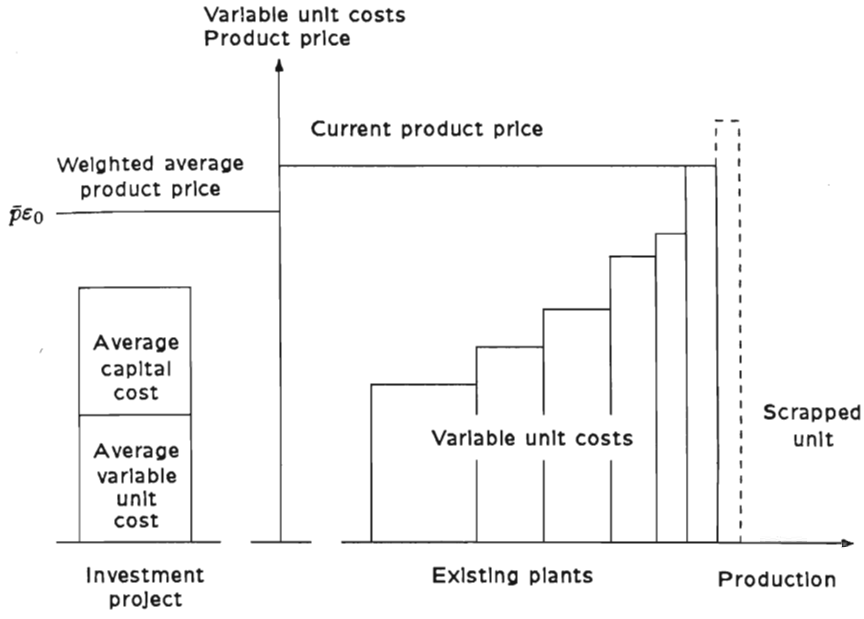
An optimal investment is characterised by the condition that the sum of current factor costs per unit produced, estimated at the average prices, plus the average capital cost per unit produced per time unit (calculated as an arithmetic average) are equal to the average product price multiplied by the elasticity of scale for the ex ante function.

As we saw above, the decision to cease operating with a vintage unit is based only on current prices. This condition and the investment condition (2.20) are illustrated in Figure 2.1.

The current output price may also be used when evaluating a new investment if the price is expected to change by a fixed percentage rate (including the value zero). If we insert (2.8) in (2.20) and consider $t = 0$, we obtain:

$$\left(\sum_{i=1}^{n}\frac{v_i}{x}q_i(0,T) + q_K(0)\frac{K}{x}\right)\cdot\frac{r-h}{1-e^{-(r-h)T}}\frac{1}{\varepsilon_0} = p(0) \tag{2.21}$$

The present value of total current factor costs and the initial factor cost per produced unit are converted to current costs per period ("yearly costs") by means of an annuity factor, with a discount rate equal to the difference between the calculation rate and the growth rate of the product price. The



Production units arranged according to increasing variable unit costs on the right-hand side, the investment project under consideration on the left-hand side.

Figure 2.1: Heckscher diagram.

yearly cost factor is multiplied by the inverse value of the elasticity of scale. With price-taking behaviour we know that $\varepsilon_0 < 1$. (It follows from (2.3) and (2.18) that we now have $\pi(0) = p(0,T) \cdot x(1-\varepsilon_0)$, so that "average yearly costs" must be less than or equal to the current product price if the investment is to fulfil the conditions for optimality in (2.7) and (2.11) When $\varepsilon_0 < 1$ the firm has a greater capital return than that corresponding to the discount rate r.

Price and demand uncertainty

Price expectation plays a focal role in the result of the vintage model, and the question naturally arises what consequences the introduction of uncertainty will have. In static analysis the problem is to maximise the expected value of a function which has profits as an argument. Allowing

for risk aversion, this function (the "utility function") is assumed to be concave.

A natural extension to our investment analysis set-up would be to link the utility evaluation at every point in time to profits. (One problem which would then arise is how to handle initial capital outlay.) The present value criterion which we used in evaluating the investment project in (2.3) is based, in general, on certain expectations.

Generally, ex ante price uncertainty leads to uncertainty concerning the utilisation of capacity at each future point in time. This is in contrast to the situation of ex ante price certainty, where one is able to determine the exact periods during which the plant will, or will not, be operated at full capacity. The two cases are generally not analogous. The firm must therefore develop a *strategy* for dealing with the effects of uncertainty at any given time.

In contrast to neoclassical static analysis, uncertainty in a vintage model implies changes in optimal full capacity inputs even in the case of risk neutrality. Earlier literature concerned with the question of flexibility of techniques in the short-run[2] has already pointed out that it is necessary to weigh expected profits (the profits that will be incurred should prices equal their expected values) against possible losses (losses that will be incurred should prices not equal their expected values): "Flexibility will be added until its 'accumulated' marginal cost equals the discounted marginal returns from savings due to that additional flexibility." (Stigler [1939, p. 316]). Furthermore, Stigler stated that one method for "securing flexibility of operations profits ... is to reduce fixed plant relative to variable services, i.e., to transform fixed into variable costs." As we shall see in a moment, this is just the general result obtained in the putty-clay model.

Price uncertainty

The consequences of price uncertainty have been studied in Kon [1983], while the case of demand uncertainty has been discussed in Albrecht and Hart [1983]. Moene [1984, 1985] looked at the simultaneous occurrance of both types of uncertainty.

In order to keep the exposition as simple as possible, the two sources of uncertainty will be treated separately. Let us start with an assumption of two time periods — the ex ante decision period and the period of actual operations. The strategic element in the decision problem is that ex post

[2] See Stigler [1939] and Lutz and Lutz [1951].

operating rules are considered when making ex ante decisions about factor proportions and capacity output. The ex post operating rule is the quasi-rent criterion: the plant is taken out of operation when the quasi-rent is negative. This criterion can be expressed as a condition on the minimal output price corresponding to (2.14):

$$p^{min}(v_1, v_2, \cdots, v_n, K) = \sum_{i=1}^{n} q_i v_i / x \tag{2.22}$$

Introducing the joint probability distribution for output and input prices,

$$h(p, q_1, q_2, \ldots, q_n) \tag{2.23}$$

the ex ante choice of the input coefficient v_i/x, $i = 1, \ldots, n$, influences the probability of shutting down the plant during the operating period. We will in this sequel assume that $Pr(p < p^{min}) > 0$.

The formal decision problem is:

$$\max E\left[pf(v_1, v_2, \cdots, v_n, K) - \sum_{i=1}^{n} q_i v_i - q_K K\right]$$

such that

$p > p^{min} \Rightarrow$ the plant is operated

$p < p^{min} \Rightarrow$ the plant is closed down

(2.23) is valid.

Equivalently, this problem can be stated as

$$\max_{v,K}\left[\int_{q=0}^{\infty}\int_{p=p^{min}}^{\infty}\left(pf(v, K) - \sum_{i=1}^{n} q_i v_i\right) h(p, q)\, dp\, dq - q_K K\right] \tag{2.24}$$

where $v = (v_1, v_2, \cdots, v_n)$, $q = (q_1, q_2, \cdots, q_n)$ and $dq = (dq_1, dq_2, \ldots, dq_n)$. When differentiating (2.24) with respect to the inputs, the lower limit p^{min} is also a function of the inputs. However, this derivative vanishes because it is evaluated at the lower limit for p, and hence the quasi-rent — the integrand in (2.24) — is zero. Differentiation of (2.24) with respect to input no i yields:

$$\int_{q=0}^{\infty}\int_{p=p^{min}}^{\infty}\left(pf_i'(v, K) - q_i\right) h(p, q)\, dp\, dq$$

Rearranging results in the following first-order condition

$$f_i'(v,K)\int_{q=0}^{\infty}\int_{p=p^{min}}^{\infty} q_i h(p,q)\,dp\,dq - \int_{q=0}^{\infty}\int_{p=p^{min}}^{\infty} ph(p,q)\,dp\,dq = 0$$

Carrying out the integrations for all input prices in the first expression and for all input prices except q_i in the second expression we obtain

$$f_i'(v,K)\int_{p=p^{min}}^{\infty} ph_p(p)\,dp = \int_{q_i=0}^{\infty}\int_{p=p^{min}}^{\infty} q_i h_{pq_i}(p,q_i)\,dp\,dq_i \qquad (2.25)$$

where $h_p(p)$ is the partial distribution for the output price and $h_{pq_i}(p,q_i)$ is the joint distribution for the output price and price of input no. i. Employing basic definitions in probability theory, (2.25) can be written:

$$f_i'(v,K)Pr(p > p^{min})E(p \mid p > p^{min}) = Pr(p > p^{min})E(q_i \mid p > p^{min})$$

or

$$f_i'(v,K)E(p \mid p > p^{min}) = E(q_i \mid p > p^{min}) \qquad (2.26)$$

In other words the certain future prices in (2.4) are replaced by expected conditional prices, the condition being that the firm will remain in business. The consequences of calculating ex ante with the possibility of not operating are seen by comparing (2.26) with the case of adjusting to the expected value of the prices

$$f_i'(v,K) = \frac{E(q_i \mid p > p^{min})}{E(p \mid p > p^{min})} < \frac{E(q_i)}{E(p)} \qquad (2.27)$$

since $E(q_i \mid p > p^{min}) < E(q_i)$ and $E(p \mid p > p^{min}) > E(p)$. The *partial* impact when facing the possibility of zero operation is to *increase* the amount of the current input in question.

By carrying out integrations similar to those from which (2.25) was derived, the differentiating of (2.24) with respect to capital, K, yields

$$f_K'(v,K)Pr(p > p^{min})E(p \mid p > p^{min}) = q_K \qquad (2.28)$$

The expected output price conditional upon full capacity utilisation times the probability of this occurrence is less than the (unconditional) expected output price, since this value can be written

$$E(p) = Pr(p > p^{min}) \cdot E(p \mid p > p^{min}) + Pr(p < p^{min}) \cdot E(p \mid p < p^{min})$$

Both expressions on the right hand side are positive. Thus we have established

$$f'_K(v,K) = \frac{q_K}{Pr(p > p^{min})E(p \mid p > p^{min})} > \frac{q_K}{E(p)} \tag{2.29}$$

The *partial* impact of an ex ante consideration of the possibility of maintaining capacity idle ex post is to *reduce* the amount of capital.

The substitution effect that results from price uncertainty is found by combining (2.26) and (2.28),

$$\frac{f'_i(v,K)}{f'_K(v,K)} = \frac{Pr(p > p^{min})E(q_i \mid p > p^{min})}{q_K} < \frac{E(q_i)}{q_K} \tag{2.30}$$

The optimal adjustment to uncertainty is to use *less* capital relative to current inputs. This is exactly the conjecture in Stigler [1939].

With respect to the substitution effect between current inputs, we have from (2.26) that the marginal rate of substitution is equal to the ratio of the *conditional* expectations of input prices. The shape of the marginal input price probability distribution now determines the deviations from the marginal rates of substitution under price certainty.

When we also consider the scale effect, i.e., the impacts on inputs of change in optimal capacity output, we see that the sign of the *total* effect has not been established. With only one current input and technical complementarity with capital, the total impact is of the same nature as the partial impacts stated above. However, with several current inputs the total effects may involve partial effects in opposite directions.

It is straightforward to extend the adjustment conditions (2.26) and (2.28) to several ex post time periods. Extending the maximisation problem in (2.24) to maximise the expected discounted profit, we obtain the problem

$$\max_{v,K}\left[\int_{t=0}^{T} e^{-rt}\int_{q(t)=0}^{\infty}\int_{p(t)=p^{min}(t)}^{\infty}\Big(p(t)f(v,K) - \sum_{i=1}^{n} q_i(t)v_i\Big)h\big(p(t),q(t)\big)\,dp(t)\,dq(t)\,dt - q_K K\right] \tag{2.31}$$

We now assume that the horizon T, that is the technical lifetime of equipment, is exogenous, and that an interior solution to (2.31) exists. More important, we also assume that prices are distributed *independently* over time.

The marginal adjustment conditions become:

$$f_i'(v,K)\int_{t=0}^{T} e^{-rt}Pr\big(p(t)>p^{min}(t)\big)\cdot E\big(p(t)\mid p(t)>p^{min}(t)\big)\,dt$$
$$=\int_{t=0}^{T} e^{-rt}Pr\big(p(t)>p^{min}(t)\big)E\big(q_i(t)\mid p(t)>p^{min}(t)\big)\,dt \quad (2.32)$$

$$f_K'(v,K)\int_{t=0}^{T} e^{-rt}Pr\big(p(t)>p^{min}(t)\big)\cdot E\big(p(t)\mid p(t)>p^{min}(t)\big)\,dt$$
$$=q_K \quad (2.33)$$

Combining (2.32) and (2.33) we see that the conclusions about substitution effects are just the same as before. Compared with the conditions (2.4) and (2.11) for the certain expectations case, we see that the price integrals defined by (2.5) are replaced by discounted present value prices expressed in terms of conditional expectations times the probabilities of operating.

Demand uncertainty

In a situation where a firm expects fluctuating demand, the ex post decision about capacity must take into consideration both the potential losses that will arise from not being able to deliver when demand exceeds capacity, and the capital costs of having excess capacity when demand is less than capacity output. The hedging against these two types of risk results in the same decrease in relative capital intensity as in the case of price and resulting quasi-rent uncertainty.

Considering again only two time-periods, the ex ante decision period and the ex post operating period, the firms's problem is to maximise the expected profit:

$$\max_{v,K} E\left[\left(pf(v,K)-\sum_{i=1}^{n} q_i v_i\right)\min\left(\frac{y}{x},1\right)-q_K K\right] \quad (2.34)$$

where y is the demand and thus $u=y/x$ the utilisation rate when $y<x$. The level of demand is the only stochastic variable and its probability

distribution is denoted $h(y)$. The problem (2.34) can be written

$$\max_{v,K}\left[\left(\int_{y=0}^{x}\frac{y}{x}h(y)\,dy+\int_{y=x}^{\infty}h(y)\,dy\right)\left(pf(v,K)-\sum_{i=1}^{n}q_iv_i\right)-q_KK\right] \tag{2.35}$$

For demand levels up to the capacity, x, the realised quasi-rent will be the full capacity rent times the rate of capacity utilisation; for demand levels exceeding capacity the quasi-rent will be the full capacity quasi-rent. Employing basic probability concepts again, (2.35) can be written as

$$\max_{v,K}\left[\left(Pr(y<x)\frac{E(y\mid y<x)}{x}+Pr(y>x)\right)\left(pf(v,K)-\sum_{i=1}^{n}q_iv_i\right)-q_KK\right] \tag{2.36}$$

The first expression is equal to the expected rate of capacity utilisation, $E(u)$. (When $y>x$, we have that $E(u\mid y>x)=1$.) The optimisation problem is therefore simply

$$\max_{v,K}\left[E(u)\left(pf(v,K)-\sum_{i=1}^{n}q_iv_i\right)-q_KK\right] \tag{2.37}$$

In Albrecht and Hart [1983] the term in front of the quasi-rent in (2.36) is called the "risk coefficient". When there is demand uncertainty, the quasi-rent at full capacity is adjusted with this risk coefficient in the ex ante decision. As we have seen, the risk coefficient is just the expected rate of capacity utilisation.

However, in order to allow the possibility of analysing consequences of increased risk in the Rothschild-Stiglitz [1970] sense of mean-preserving spread, Albrecht and Hart express the risk-coefficient in terms of the distribution function, $H(\cdot)$, for demand y. Integrating the first integral in (2.35)

by parts, the risk coefficient becomes:

$$\begin{aligned} E(u) &= \frac{1}{x}\int_0^x yh(y)\,dy + \int_x^\infty h(y)\,dy \\ &= \frac{1}{x}\left(\Big|_0^x yH(y) - \int_0^x H(y)\,dy\right) + \big(1 - H(x)\big) \\ &= H(x) - \frac{1}{x}\int_0^x H(y)\,dy + \big(1 - H(x)\big) \\ &= 1 - \frac{1}{x}\int_0^x H(y)\,dy \end{aligned} \tag{2.38}$$

"Fatter tails" to the distribution function $H(y)$ for the same mean implies a *smaller* value of the risk coefficient. However, we will not pursue the analysis of riskier demand here, but rather limit ourselves to deriving the marginal conditions corresponding to price uncertainty. Differentiating (2.37) with respect to input no. i yields:

$$E(u)\big(pf_i'(v,K) - q_i\big) + \frac{\partial E(u)}{\partial v_i}\big(pf_i'(v,K) - \sum_{i=1}^{n} q_i v_i\big) \tag{2.39}$$

When deriving the expression for $\partial E(u)/\partial v_i$, it must be remembered that the lower and upper limit of the respective integrals in the expression for $E(u)$ are functions of the inputs, and hence yield

$$\begin{aligned} \frac{\partial E(u)}{\partial v_i} &= \frac{1}{x}xh(x)f_i'(v,K) - \frac{1}{x^2}f_i'(v,K)\int_0^x yh(y)\,dy - h(x)f_i'(v,K) \\ &= -\frac{1}{x^2}f_i'(v,K)\int_0^x yh(y)\,dy \\ &= -\frac{1}{x}f_i'(v,K)Pr(y<x)E(u \mid y<x) \end{aligned} \tag{2.40}$$

When differentiating $E(u)$ with respect to capital the same expression results with $f_K'(\cdot)$, replacing $f_i'(\cdot)$ in (2.40). Inserting (2.40) into (2.39) and rearranging yields the marginal adjustment conditions

$$f_i'(v,K)P = q_i E(u) \tag{2.41}$$

$$f_K'(v,K)P = q_K \tag{2.42}$$

where

$$P = \left[pE(u) - \frac{1}{x} Pr(y < x) E(u \mid y < x) \left(px - \sum_{i=1}^{n} q_i v_i \right) \right]$$

The downward adjustment of the product price due to demand uncertainty involves both a general multiplicative correction by the expected rate of capacity utilisation, and a (negative) additive term for the demand constrained case involving the conditional expectation of the unit quasi-rent.

Combining (2.41) and (2.42) gives the marginal rate of substitution between a current input and capital

$$\frac{f_i'(v,K)}{f_K'(v,K)} = \frac{q_i E(u)}{q_K} \tag{2.43}$$

Compared to a situation without demand uncertainty a *less* capital intensive technique is chosen, again supporting Stigler's conjecture.[3] As to the substitution effects between current inputs, note from (2.41) that the marginal rate of substitution is the same as in the case of certain demand. The adjustment factors for demand uncertainty are *input neutral.* They are the same, independent of the input in question.

From (2.41) the *partial* effect on the optimal amount of a current input is seen to be negative. The relative reduction in output price is greater than the reduction in input price. In the case of capital demand it is only the output price that is reduced, thus generating a negative effect on demand. As to the total effect, one way of establishing a comparison is to see in which case the capacity chosen is larger — when demand is certain or when demand is uncertain. We will not pursue such a comparison here.

Assuming a given horizon, T, and that demand is distributed independently over time, the extension of the analysis to several time periods is straightforward. Problem (2.34) can be written

$$\begin{aligned} \max_{v,K} E\Bigg[\int_{t=0}^{T} e^{-rt} \Bigg(\Bigg(p(t) f(v,K) - \sum_{i=1}^{n} q_i(t) v_i \Bigg) \\ \cdot \min \left(\frac{y(t)}{x}, 1 \right) \Bigg) dt - q_K K \Bigg] \end{aligned} \tag{2.44}$$

[3] See Stigler [1939].

The marginal conditions (2.41) and (2.42) become:

$$f_i'(v,K)P(0,T) = \int_{t=0}^{T} e^{-rt} q_i(t) E\big(u(t)\big)\, dt \tag{2.45}$$

$$f_K'(v,K)P(0,T) = q_K \tag{2.46}$$

where

$$P(0,T) = \int_{t=0}^{T} e^{-rt}\Big[p(t)E\big(u(t)\big) - \frac{1}{x} Pr\big(y(t) < x\big) E\big(u(t) \mid y(t) < x\big) \cdot \Big(p(t)x - \sum_{i=1}^{n} q_i(t) v_i\Big)\Big]\, dt$$

The substitution effects are of the same nature as those discussed for one ex post period with *present value* of the prices weighted with "risk coefficients" replacing the one period prices.

2.3 The notion of optimal structure and optimal structural change

Introduction

The ex ante – ex post framework and the vintage model presented in the last section serves as a point of reference for the conceptual discussion in this section. The discussion closely refers to the industrial policy debate in Scandinavia, as commented upon in Section 1.3.

In Scandinavia an important part of the debate on economic policy is dealing with the structural efficiency of various industries and the policy measures that should be taken to promote a more rational structure of certain industries. Interest has chiefly focused on the modernity of capital equipment, the size of plants and the extent of division of labour and specialisation. Considerable research in this area has been concerned with the number of mergers and cooperation agreements occurring within an industry over various time periods. Government commissions of inquiry have been appointed to survey the structure of different industries and to recommend measures which, so it is believed, will accelerate the process leading to a more *efficient* structure. The fundamental concepts in the debate, albeit extremely vague and imprecise, have been *optimal structure* (rational structure) and *optimal structural change* (structural rationalisation).

Consequently, *structural rationalisation*, *structural transformation*, *rational structure* and the like are words frequently used in the Scandinavian debate on economic policy. However, the meaning of these terms is most ambiguous and they have never been given a satisfactory theoretical treatment. A comparative-static framework has dominated the discussion. This is probably due in part to the difficulties of anchoring the concepts in traditional production theory.

The purpose of this section is to explain and define the terms cited above. The main point is that the terms must be analysed dynamically and not statically.

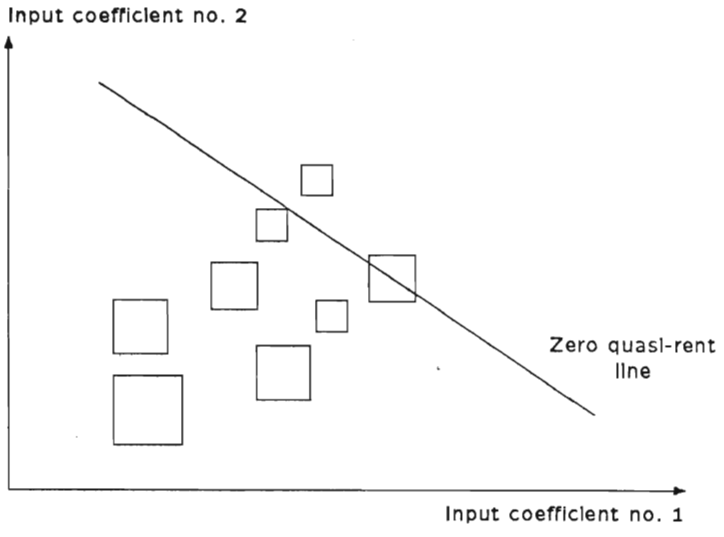
Optimal structure and optimal structural change

In an industry where investment decisions are taken within an environment of changing prices and technology, one finds at each point in time a specific distribution of capacities and input coefficients for the micro units. This was illustrated in Figure 2.1. The distribution may in the two factor case be conveniently described by a diagram in the input-coefficient space as in Figure 2.2, where the capacity of each micro unit is also indicated graphically by the size of the squares.

In Section 2.2 it was shown that a nonnegative quasi-rent was required for operation of a micro unit. This condition may also be introduced in the capacity distribution diagram by entering the line where the quasi-rent is zero (see Figure 2.2). In general the iso-quasi-rent lines correspond to the isocost lines for current inputs in the input space, i.e., their slopes are equal to the factor price ratio.

Every change which occurs in the capacity distribution may be characterised as a *structural transformation*, *structural change* or *structural development*. These three terms, which are purely descriptive, are regarded as synonymous. "Optimal structure" and "optimal structural change" on the other hand have to do with economic efficiency.

The use of the concept *optimal structure* in the Scandinavian debate leaves the impression that optimal structure is defined as the cost-minimising production structure for the case where all existing productive equipment is scrapped and only the newest technology employed. This would mean (under certain assumptions) that the capacity distribution consisted of a single point in the capacity distribution diagram, embracing the whole industry's capacity. In this sense the concept envisions a future equilibrium situation where the plants are all of cost-minimising size and of identical technique.



This figure comprises the micro units shown in Figure 2.1

Figure 2.2: Capacity distribution diagram.

This definition resembles the long-run production function for an industry, a more hypothetical construct which is closely associated with the ex ante function.[4] In the long-run industry production function it is assumed that at any given time there is a certain amount of capital and current inputs for the industry as a whole. Furthermore, it is hypothesised that capital is malleable and can take on any form desired. Under these conditions total output is maximised as a function of capital and current inputs. How is this function related to the current production possibilities? Only in the case where all capacity is concentrated at a single input coefficient-point can a point on the long-run production function be realised.

It is only in a stationary state that the *static optimal structure* concept, implied in the debate on structural rationalisation, is relevant. The term becomes misleading in a dynamic analysis of an industry in which technology is advancing and prices are changing. In this situation the "optimal structure" then changes constantly. If we optimise over a longer time interval, the "optimal structure" is never optimal. A broad scatter of the units in the input-coefficient diagram or the capacity distribution cannot

[4] See Section 1.4.

directly lead to the conclusion that the structure is nonoptimal. On the contrary, the more rapid the rate of advancing technology in an industry, the greater should be the differences that arise between the oldest and newest capital vintages. Structural change is a completely normal development in a dynamic economy. A scattered structure cannot be regarded as nonoptimal. As long as the units have nonnegative quasi-rents, they have their *raison d'être*, as we saw in Section 2.2.

This frequently used definition of optimal structure leads the debate in the wrong direction. It gives a deceptive appearance of perpetual dissatisfaction with the existing structure, a dissatisfaction that has no basis from a dynamic perspective. On the other hand, the definition does reflect a not unusual, comparative static line of thinking, with its aspiration towards an equilibrium (i.e., a single point on the long-run industry production function) where the structure consists solely of new modern units. However, it must not be forgotten that the process of structural change also entails costs. The concept of static optimal structure should be dropped in favour of what we shall call below *best-practice structure.*

The problem is not to bring the existing structure closer to the best-practice structure or to a certain state, but rather to optimise a process that is going on all the time. The existing structure, taken together with new choices of technique from the ex ante function, gives rise to a continuous structural transformation.

Figure 2.1 in Section 2.2 provides a snapshot of a structural development with respect to the following structural variables: current unit costs, quasi-rent per produced unit and market shares. The capacity distribution diagram in Figure 2.2 provides another snapshot of structural development.

Structural development can be further illustrated by means of another diagram of this kind. In Figure 2.3 the capacity distributions for two different years t_1 and t_2 are shown. Owing to technical progress the capacity distribution for the most recent year, t_2, has moved towards the origin, with lower input requirements for best-practice plants. The zero quasi-rent line has also moved during the time between year t_1 and t_2 on account of changing prices. Units with negative quasi-rent in year t_1 are not shown. Some of the units at t_2 may be completely new with the choice-of-technique based on the ex ante function, and some may be modernised units which existed at t_1. Without any changes in input coefficients only two units with the t_1-technology would earn a positive quasi-rent at t_2.

An optimality concept, such as optimal structural development, is generally based on the solution of an optimisation problem. Such an optimisation problem may be faced by an industrial organisation or a government

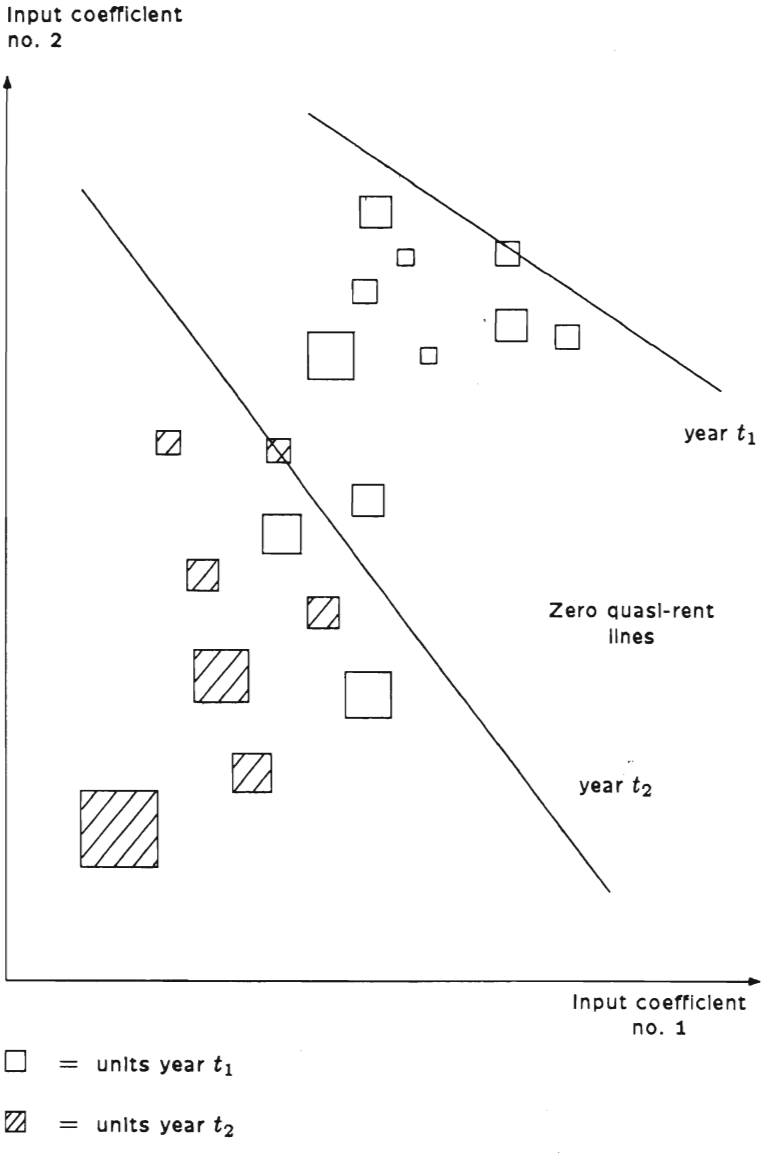


Figure 2.3: Structural development illustrated by a change in the capacity distribution.

agency in charge of industrial policy. If *free competition* is the predominant institutional framework of the economy, the objective of the agencies in question is probably only to secure that the market mechanism functions as smoothly as possible.

Optimal structural development is defined as development where the investment criteria in Section 2.2, Equations (2.4), (2.11) and (2.14), and the quasi-rent scrapping criteria are fulfilled. *Structural rationalisation* is then defined as the measures necessary for the fulfilment of these conditions.

In a dynamic perspective, however, the market mechanism may have some inherent shortcomings. How is the correct coordination of total supply and demand in the future achieved when the individual units regard prices as exogenous and foresee no problems in acquiring inputs or selling their products? Individual units may have uniform or different price expectations, but in either case expectations may be wrong. Such discrepancies are not revealed in a current market equilibrium.[5]

When a governmental agency pursues a more active industrial policy than merely facilitating the functioning of the market economy, we may regard the institutional set up as a *mixed economy.* In this case, the authorities may compute a cost-minimising structural development for industry based on forecasts about future techniques, demand functions and prices. At each moment in time this optimal structural development will be conditional upon the forecasts, but the investment and scrapping decisions are still made on a decentralised basis. *Structural rationalisation* is now defined as the measures employed to influence the decentralised investment and scrapping-decisions in order to ensure that current decisions conform with the optimal structural development valid at each point in time.

Coordination is a real problem in a dynamic setting with decentralised units because, as mentioned above, each micro unit does not foresee any *future* problems in acquiring planned inputs at forecasted prices. In a mixed economy, moreover, it is most reasonable to assume that planning agencies can only forecast prices rather than predetermine prices.

In a *centrally planned economy* the number of exogenous price forecasts is reduced to a minimum. As an extreme case, we may envisage the economy running according to a predetermined plan. The prices appear as shadow prices in the solution of the planning problem.[6] Optimal structural development is determined by the plan. More realistically, some of the prices may be inherently exogenous, such as prices of imported and ex-

[5] See Johansen [1967].

[6] *ibid.*

ported goods. Technological development to a large extent also constitutes an exogenous variable.

2.4 Economies of scale and optimal capacity expansion in a putty-clay model

Introduction

The purpose of this section is to look more closely at the optimal path of capacity expansion for an industry under putty-clay assumptions and economies of scale. There is abundant evidence of considerable scale economies, ex ante, in most manufacturing industries.[7] Under static assumptions one obtains the result that it is optimal to have one big single plant. When demand grows for the industry's product, an optimal process means that it is optimal at each point in time to have several plants, even if the ex ante function is constant over time and no technological progress occurs. Capacity needs to increase and a trade-off problem arises between the unwanted overcapacity at the beginning of the period and the favourable exploiting of economies of scale. This point is elaborated in Section 2.4. The main objective of this analysis is to further our understanding of the nature of the structural transformation process within industries. In particular we address two old topics, namely, the size distribution of plants in different industries and the trade-off problem between economies of scale and monopoly power. We purport to show that the model presented in this section sheds new light on these topics.

The vintage model presented in Section 2.2 is based on a heterogeneous ex ante production function. In order to focus more explicitly on economies of scale the model presented in this section contains a homogeneous ex ante production function with a scale elasticity exceeding 1. Confronted with empirical evidence, a dynamic model based on increasing returns to scale and large substitution possibilities in the ex ante production function but with fixed maximum capacity and frozen factor proportions ex post is easy to defend.

[7] See, e.g., Scherer et al. [1975], Pratten [1971] and Haldi and Whitcomb [1967].

The importance of economies of scale

According to empirical studies, economies of scale in the ex ante production function seem to be important for most manufacturing industries.[8] In Table 2.1 a summary of empirical estimates of economies of scale is presented. Assuming a constant scale elasticity the estimates hold for the shown range of capacity.[9]

In Scherer et al. (1975, Chapter 3), the minimum optimal scales, MOS, of different industries in 1967 are compared to domestic consumption, where MOS is the smallest capacity or planned output volume at which all relevant economies of scale are achieved. As can be seen in the Table 2.2 there is a clear difference between Sweden and the other countries.

The trade-off problem between scale efficiency and market power seems to be a small one in the US and other large countries, but important to small countries like those of Scandinavia. These latter countries do not seem to worry very much about the number of domestic firms in an industry, due to reliance on world market competition. Their industrial policy also seems to stress scale efficiency at the expense of *domestic* competition.

An important question is the strategic behaviour of the firms in industries characterised by scale economies. Without referring to any systematic empirical investigation a number of casual empirical observations support the view that investment decisions are often based on the expected growth in output demand and goals about market shares, together with cost calculations and expected rate of return or simple pay-off criteria, rather than by a maximisation of an explicit discounted profit function as in Equation (2.3). Thus, the behaviour may be characterised as cost-minimising and "demand-taking" instead of price-taking.

Interrelated issues concerning the cost-minimisation problem raised by the existence of economies of scale are the optimal timing of investments and how much excess capacity to permit when a new plant is to come on line. It is possible to decrease average unit cost a certain amount by overbuilding when plant construction exhibits increasing returns.

Several applied models assuming economies of scale have been constructed in a planning framework to focus on optimal plant size and construction timing in single industries. These models are primarily concerned with the trade-off between the cost of having excess capacity and the lower

[8] See Pratten (1971) for a thorough study.

[9] For further details, see Hjalmarsson (1976).

Table 2.1: Estimates of economies of scale.

Product, etc.	Source of data	Range of physical capacity	Elasticity of scale: ε
Refinery	Ribrant p. 251	(4–6) ktons p.a.	1.35
Ethylene plants	Ribrant p. 265	(100–300) ktons p.a.	1.17
Ethylene plants	Ribrant p. 265	(50–200) ktons p.a.	1.24
Sulphuric acid plants	Pratten p. 50	(100–1000) ktons p.a.	1.03
Dye plants	Pratten p. 52	(0.75–1.5) ktons p.a.	1.40
Polymer plants	Pratten p. 65	(4–80) ktons p.a.	1.07
Polymer plants	Pratten p. 65	(20–80) ktons p.a.	1.07
Beer breweries	Pratten p. 74	(0.1–1.0) million barrels p.a.	1.24
Beer breweries	Pratten p. 74	(0.2–1.0) million barrels p.a.	1.25
Bread bakeries	Ribrant p. 352	(0.9–1.8) tons p.h.	1.27
Sugar refinery plants	Ribrant p. 360	(1.1–4.2) tons p. 24 hs.	1.09
Milk dairies	Ribrant p. 370	(10–40) ktons p.a.	1.55

Table 2.1: Estimates of economies of scale, continued.

Product, etc.	Source of data	Range of physical capacity	Elasticity of scale: ε
Butcheries	Ribrant p. 380	(2–8) ktons p.a.	1.21
Butcheries	Ribrant p. 380	(2–4) ktons p.a.	1.40
Detergent plants	Pratten p. 86	(10–70) ktons p.a.	1.05
Detergent plants	Pratten p. 86	(30–70) ktons p.a.	1.03
Cement portland	Pratten p. 92	(0.1–2.0) million tons p.a.	1.18
Cement works	Ribrant p. 209	(0.12–1.0) million tons p.a.	1.38
Crude steel plants	Pratten p. 15	(0.25–10) million tons p.a.	1.09
Steel blast furnaces	Pratten p. 106	(265–400) ktons p.a.	1.34–1.93
Pulp plants	Wohlin p. 77	(67–268) ktons p.a.	1.28
Newspaper pulp plants	Wohlin p. 77	(55–440) ktons p.a.	1.19

unit-capacity costs permitted by overbuilding. Manne and his associates[10] applied a constant elasticity capacity cost function in a model for the planning of capacity expansion in India. Indeed, the model we set out here has certain similarities with the models of Manne and Srinivasan, and can be seen as a further development of these. On closer inspection it turns out

[10] See Manne et al. [1967].

Table 2.2: The number of MOS plants compatible with domestic consumption in six nations, 1967

	Nation					
Industry	U.S	Canada	U.K.	Sweden	France	Germany
Brewing	29.0	2.9	10.9	0.7	4.5	16.1
Cigarettes	15.2	1.3	3.3	0.3	1.6	2.8
Fabrics	451.7	17.4	57.0	10.4	56.9	52.1
Paints	69.8	6.3	9.8	2.0	6.6	8.4
Petroleum refining	51.6	6.0	8.6	2.5	7.7	9.9
Shoes	532.0	59.2	164.5	23.0	128.2	196.9
Glass bottles	65.5	7.2	11.1	1.7	6.6	7.9
Cement	59.0	6.6	16.5	3.5	21.7	28.8
Steel	38.9	2.6	6.5	1.5	5.5	10.1
Bearings	72.0	5.9	22.8	3.3	17.0	n.a
Refrigerators	7.1	0.7	1.2	0.5	1.7	2.8
Storage batteries	53.2	4.6	7.7	1.4	12.8	10.5

Source: Scherer et al. (1975, p. 94).

that the Srinivasan model is a special case of our model. The Srinivasan model is based on a simple cost function whereas our model presupposes a production function with ex ante substitution possibilities between two current factors and capital. Our main purpose is not to contribute with a novel model of capacity expansion, but to study the distribution of capacity and input coefficients generated by such a model.

The model

The problem can be formulated as follows: An industry produces a homogeneous product. What is the optimal sequence for the time of construction and what are the optimal size of plants in order that domestic production entirely satisfy demand at each future point in time? We assume the following:

(i) Demand grows at a constant exponential rate g.

(ii) Initially there is just enough capacity, denoted by $x(0,0) = x$, to meet the demand.

(iii) The ex ante function at the micro level exhibits increasing returns to scale and is a quasi-concave function with capital equipment and two current inputs. For the sake of mathematical and computational simplicity we choose a Cobb-Douglas with neutral technological change. The ex ante function (2.1) now reads

$$x(\nu,\nu) = Ae^{\delta\nu}L(\nu,\nu)^{a_L}E(\nu,\nu)^{a_E}K(\nu,\nu)^{a_K} \tag{2.47}$$

where $a_L + a_E + a_K = \varepsilon > 1$ and where $x(\nu,\nu)$, $L(\nu,\nu)$, $E(\nu,\nu)$ and $K(\nu,\nu)$ are planned production at time ν with vintage ν, planned use of variable inputs and planned capital investment, respectively. δ is the technical progress parameter.

(iv) The following functions describe the change in the factor prices:

$$q_i(t) = q_i(0)e^{\gamma_i t} \qquad i = L,\ ,E,\ K \tag{2.48}$$

where $q_i(0)$ is the initial price.

(v) Plant life is infinite and the time horizon is infinite.

(vi) Capacity utilisation in the most recent plant grows at the same rate as demand until the next investment point, at which there is no unutilised capacity, is reached. The assumption is partly made for convenience and partly based on the following consideration: if the time period between two investments is not too long we may regard it as an initial adjustment period. During this period utilisation of capacity grows continuously.

An alternative approach to assumption (vi) is to allow during the initial stages full capacity utilisation in the most recent plant, and let the capacity utilisation vary in the oldest plants with the highest unit costs. As time goes on, more and more of the older plants would be involved in this process of fluctuating capacity utilisation. Typically, however, one must account for inertia and some costs in restarting old equipment. If these costs are considerable, optimisation procedures would show that it is sometimes more advantageous to build a new plant which is somewhat larger than demand, so as to avoid having to start an old plant again. Even without inertia, the same thing would happen, in part because of embodied technological change, in part because the existence of economies of scale makes it more profitable to build a somewhat larger plant with low-unit costs than using

the oldest plants with high-unit costs. The putty-clay assumption together with disproportional price developments tend to produce similar effects. Thus the smallest plants continuously disappear and a size distribution results with fewer and somewhat larger plants.
Concerning cost minimisation these problems are not too serious, since discounting results in the first years having the most influence over the investment decision. What happens to a plant in the remote future when it constitutes only an insignificant part of total capacity (even if it is a "big" plant when it is erected) is of little importance to the determination of total costs over the whole time horizon. A priori, both assumptions suggested above are possible. Which is more realistic is an empirical question. We have choosen the first assumption, (vi), for the sake of mathematical simplicity.

(vii) Discrete time periods are assumed. To distinguish between the different vintages, successive time points of investments are denoted by τ_n ($\tau_0 = 0$, $n = 0, 1, 2, \ldots,$) which in general may differ from the index for real time. It is assumed that the first time point of investment coincides with the starting point zero. During the interval between two successive installations there are at least two possibilities for the amount of input required during the time when capacity utilisation in the most recent plant grows at the same rate as demand. Input coefficients may be fixed at the full capacity level independent of capacity utilisation or they may decrease when the rate of capacity utilisation increases. The former assumption is adopted here while the latter assumption is also considered in Hjalmarsson[1974].

The assumptions (i)–(vii) above imply the following important constant cycle time theorem:

Theorem 2.1: An optimal policy consists of building successive plants at equidistant intervals of time.

Proof: See Appendix 2.1.

The time interval between two investment points is denoted by τ and $\tau_n = n\tau$, $n = 0, 1, 2, \ldots$.

The growth in demand during the interval τ_n to τ_{n+1} is

$$xe^{g\tau_{n+1}} - xe^{g\tau_n} = xe^{ng\tau}(e^{g\tau} - 1) \tag{2.49}$$

This expression must be equal to the capacity installed at time τ_n

$$Ae^{\delta\tau_n}\overline{L}(\tau_n,\tau_n)^{a_L}\overline{E}(\tau_n,\tau_n)^{a_E}\overline{K}(\tau_n,\tau_n)^{a_K} \tag{2.50}$$

where the bars indicate full capacity values.

The cost of the plant to be constructed at time point τ_n, discounted to year 0, is denoted by C_{τ_n} and is given by the expression

$$\begin{aligned} C_{\tau_n} = &\left(\sum_{t=\tau_n}^{\infty} q_L(t)e^{-rt}\cdot\overline{L}(t,\tau_n)\right)+\left(\sum_{t=\tau_n}^{\infty} q_E(t)e^{-rt}\cdot\overline{E}(t,\tau_n)\right)\\ &+q_K(\tau_n)e^{-r\tau_n}\cdot\overline{K}(\tau_n,\tau_n)\end{aligned} \tag{2.51}$$

The expression is to be minimised under the constraint that capacity (2.50) equals demand (2.49). We then obtain the following first-order condition:

$$\begin{aligned}\frac{\sum_{t=\tau_n}^{\infty} q_L(t)e^{-rt}\cdot\overline{L}(t,\tau_n)}{a_L} &= \frac{\sum_{t=\tau_n}^{\infty} q_E(t)e^{-rt}\cdot\overline{E}(t,\tau_n)}{a_E}\\ &= \frac{q_K(\tau_n)e^{-r\tau_n}\cdot\overline{K}(\tau_n,\tau_n)}{a_K}\end{aligned} \tag{2.52}$$

$$Ae^{\delta\tau}\cdot\overline{L}(\tau_n,\tau_n)^{a_L}\cdot\overline{E}(\tau_n,\tau_n)^{a_E}\cdot\overline{K}(\tau_n,\tau_n)^{a_K} = xe^{ng\tau}(e^{g\tau}-1) \tag{2.53}$$

The conditional factor demand functions are derived from Equations (2.52) and (2.53) by inserting (2.48) which yields

$$\begin{aligned}\overline{L}(\tau_n,\tau_n)^{\varepsilon} = xA^{-1}\cdot a_L^{\varepsilon}\prod_i a_i^{-a_i}\left(1-e^{\gamma_K-r}\right)^{a_K}\left(\frac{q_L(0)}{1-e^{\gamma_L-r}}\right)^{-\varepsilon}\\ \prod_i\left(\frac{q_i(0)}{1-e^{\gamma_i-r}}\right)^{a_i}(e^{g\tau}-1)\cdot e^{\left(\sum_i\gamma_i a_i-\gamma_L\varepsilon-\delta+g\right)n\tau}\end{aligned} \tag{2.54}$$

$$i = L,\ E,\ K$$

$$\begin{aligned}\overline{E}(\tau_n,\tau_n)^{\varepsilon} = xA^{-1}\cdot a_E^{\varepsilon}\prod_i a_i^{-a_i}\left(1-e^{\gamma_K-r}\right)^{a_K}\left(\frac{q_E(0)}{1-e^{\gamma_E-r}}\right)^{-\varepsilon}\\ \prod_i\left(\frac{q_i(0)}{1-e^{\gamma_i-r}}\right)^{a_i}(e^{g\tau}-1)\cdot e^{\left(\sum_i\gamma_i a_i-\gamma_E\varepsilon-\delta+g\right)n\tau}\end{aligned} \tag{2.55}$$

$$i = L,\ E,\ K$$

$$\overline{K}(\tau_n,\tau_n)^{\varepsilon} = xA^{-1}\cdot a_K^{\varepsilon}\prod_i a_i^{-a_i}\left(1-e^{\gamma_K-r}\right)^{a_K} q_K(0)^{-\varepsilon}$$
$$\prod_i\left(\frac{q_i(0)}{1-e^{\gamma_i-r}}\right)^{a_i}(e^{g\tau}-1)\cdot e^{(\sum_i \gamma_i a_i-\gamma_K\varepsilon-\delta+g)n\tau} \quad (2.56)$$
$$i = L,\ E,\ K$$

From Equation (2.51) together with (2.53)–(2.56) we obtain the following cost function:

$$C_{\tau_n} = B\left(e^{g(\tau_{n+1}-\tau_n)}-1\right)^{\frac{1}{\varepsilon}} e^{\gamma\tau_n} \quad (2.57)$$

where

$$B = H^{1/\varepsilon}\cdot x^{1/\varepsilon}\cdot\varepsilon \quad (2.58)$$

$$H = A^{-1}\cdot\prod_i a_i^{-a_i}\left(1-e^{\gamma_K-r}\right)^{a_K}\prod_i\left(\frac{q_i(0)}{1-e^{\gamma_i-r}}\right)^{a_i} \quad (2.59)$$
$$i = L,\ E,\ K$$

$$\gamma = \frac{\sum_i \gamma_i a_i-\delta+g}{\varepsilon}-r \qquad i = L,\ E,\ K \quad (2.60)$$

From the constant cycle time theorem (Theorem 2.1) we have $\tau_n = n\tau$.

Summation over all n yields the total cost function for the whole horizon as a function of the time interval. It is denoted by $C(\tau)$ and includes the discounted stream of construction costs as well as operation costs:

$$C(\tau) = \sum_{n=0}^{\infty} C_{\tau_n} = B\cdot\frac{(e^{g\tau}-1)^{1/\varepsilon}}{1-e^{\gamma\tau}} \quad (2.61)$$

where $\gamma < 0$ and $B > 0$.

If $\gamma > 0$, C_{τ_n} in (2.57) is strictly increasing and $\sum_{n=0}^{\infty} C_{\tau_n}$ does not converge. We have $C(\tau)\to\infty$ for $\tau\to 0$, but a minimum may not exist for all parameter values.[11]

The optimal time interval is obtained by minimising $C(\tau)$ with respect to τ. Differentiating $\log C(\tau)$ with respect to τ and equating the derivate with

[11] See Appendix 2.2.

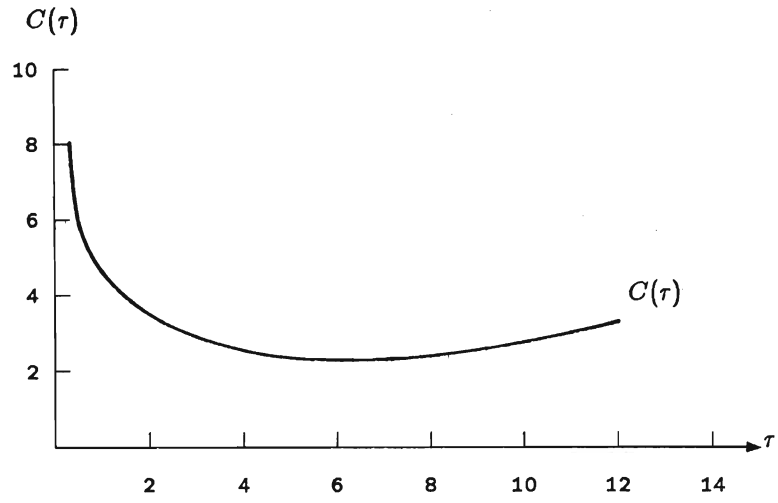


Figure 2.4: The cost function for the parameter values $B = 1$, $\gamma = -0.053$, $\varepsilon = 1.50$, $\gamma_i = \delta = 0$, $g = 0.10$ and $r = 0.12$.

zero we get the following first-order condition:

$$\frac{C'(\tau)}{C(\tau)} = \frac{1}{\varepsilon} \cdot \frac{ge^{g\tau}}{e^{g\tau}-1} - \frac{\gamma e^{\gamma\tau}}{e^{\gamma\tau}-1} = 0 \tag{2.62}$$

In Figure 2.4 we have calculated the function for specific values of the parameters. In Appendix 2.2 it is shown that $C(\tau)$ has a unique minimum.

For a high value of g together with high values of γ, i.e., low absolute values, and low values of ε, the neighbourhood around the minimum point becomes more curved. If g is low, however, the curves are rather flat independently of γ and ε. The Tables 2.3, 2.4 and 2.5 below show the optimal time cycle for various values of the parameters γ, ε and g.

From the expression for capacity increment $xe^{ng\tau}(e^{g\tau}-1)$ and $x = 100$ we have calculated the optimal capacity expansion for different time periods and values of the parameters. The capacity distribution in the input-coefficient space is illustrated in the tables and Figure 5 below. The size distribution of plants generated by the process of optimal capacity expansion is further analysed in Section 2.5, as is scale efficiency and a comparison of the costs of a decentralised versus a centralised capacity expansion in Section 2.6.

Table 2.3: The value of τ for $g = 0.10$.

γ \ ε	1.10	1.25	1.50	1.75	2.00	2.25	2.50
−0.01	1.76	4.19	7.83	11.09	14.07	16.82	19.50
−0.02	1.61	3.82	7.10	9.99	12.60	14.99	17.20
−0.03	1.48	3.51	6.49	9.09	11.42	13.53	15.47
−0.04	1.38	3.25	5.97	8.34	10.45	12.34	14.06
−0.05	1.28	3.02	5.54	7.71	9.63	11.35	12.90
−0.06	1.20	2.82	5.16	7.17	8.93	10.50	11.92
−0.07	1.13	2.65	4.83	6.70	8.33	9.78	11.09
−0.08	1.07	2.49	4.54	6.29	7.81	9.15	10.36

Table 2.4: The value of τ for $\gamma_i = \delta = 0$ $(i = L, E, K)$ and $g = r = 0.10$.

ε	1.10	1.25	1.50	2.00
τ	1.95	3.82	6.48	9.63

Table 2.5: The value of τ for $\gamma_i = \delta = 0$ $(i = L, E, K)$, $\varepsilon = 1.25$ and $r = 0.10$.

g	0.03	0.05	0.08	0.10	0.12
τ	4.15	4.05	3.91	3.82	3.74

The capacity distribution in the input coefficient space

Since our main interest is in the choice of factor proportions and the development of industrial structure we have to look more closely at the distribution of input coefficients.

Input per unit of output, the input coefficient for vintage τ_n is denoted by $\xi_i(\tau_n)$. It is a variable ex ante, but a fixed coefficient ex post. From Equations (2.54)–(2.56) and (2.49) one obtains

$$\xi_L(\tau_n) = \frac{\overline{L}(\tau_n, \tau_n)}{\bar{x}(\tau_n, \tau_n)} = A_L \cdot (e^{g\tau} - 1)^{(1-\varepsilon)/\varepsilon} \cdot e^{Bn\tau} \tag{2.63}$$

where

$$A_L = x^{(1-\varepsilon)/\varepsilon} \cdot a_L \left(\frac{q_L(0)}{1 - e^{\gamma_L - \tau}} \right)^{-1} \cdot H^{1/\varepsilon} \tag{2.64}$$

$$B = \left(\sum_i \gamma_i a_i - \gamma_L \varepsilon - \delta + (1-\varepsilon) g \right) / \varepsilon \tag{2.65}$$

and

$$\xi_E(\tau_n) = \frac{\overline{E}(\tau_n, \tau_n)}{\bar{x}(\tau_n, \tau_n)} = A_E \cdot (e^{g\tau} - 1)^{(1-\varepsilon)/\varepsilon} \cdot e^{Cn\tau} \tag{2.66}$$

where

$$A_E = x^{(1-\varepsilon)/\varepsilon} \cdot a_E \left(\frac{q_E(0)}{1 - e^{\gamma_E - \tau}} \right)^{-1} \cdot H^{1/\varepsilon} \tag{2.67}$$

$$C = \left(\sum_i \gamma_i a_i - \gamma_E \varepsilon - \delta + (1-\varepsilon) g \right) / \varepsilon \tag{2.68}$$

and

$$\xi_K(\tau_n) = \frac{\overline{K}(\tau_n, \tau_n)}{\bar{x}(\tau_n, \tau_n)} = A_K \cdot (e^{g\tau} - 1)^{(1-\varepsilon)/\varepsilon} \cdot e^{Dn\tau} \tag{2.69}$$

where

$$A_K = x^{(1-\varepsilon)/\varepsilon} \cdot a_K \cdot q_K(0)^{-1} \cdot H^{1/\varepsilon} \tag{2.70}$$

$$D = \left(\sum_i \gamma_i a_i - \gamma_K \varepsilon - \delta + (1-\varepsilon) g \right) / \varepsilon \tag{2.71}$$

The level of the input coefficients is determined by a complicated formula involving several parameters and it is not easy to distinguish the relative significance of different parameters. The relative importance of the parameters determining the time path of the input coefficients is easier to grasp.

Note the importance of the sign of the own factor price development and the interaction between economies of scale and demand growth. Note also that the time path — but not the level of the input coefficient — is independent of the rate of interest.

The development of the ratio between two input coefficients when new capacity is built is given by

$$\frac{\xi_i(\tau_n)}{\xi_j(\tau_n)} = D_{ij}e^{(\gamma_j-\gamma_i)n\tau} \qquad i,\ j = L,\ E,\ K \tag{2.72}$$

where

$$D_{ij} = \frac{a_i}{a_j} \cdot \frac{q_j(0)}{q_i(0)} \cdot \frac{1-e^{\gamma_i-r}}{1-e^{\gamma_j-r}} \qquad i,\ j = L,\ E \tag{2.73}$$

$$D_{iK} = \frac{a_i}{a_K} \cdot \frac{q_K(0)}{q_i(0)} \cdot (1-e^{\gamma_i-r}) \qquad i = L,\ E \tag{2.74}$$

i.e., the development of the relative factor ratio is only governed by the difference in factor price change between the two inputs. If the factor prices change at the same rate, $\gamma_i = \gamma_j$, the factor ratio is constant.[12]

The constant term is also fairly simple. Obviously a large marginal elasticity and a high initial own factor price contribute to a large factor ratio. It is reasonable to assume that in most cases the rate of interest is higher than the factor price change. Then a rapid increase in the price of a current factor γ_i reduces the ratio.

Assuming $x = 100$, $A_0 = 4$, $q_L(0) = 10$, and $q_K(0) = 1$, input coefficients and capacity have been calculated for different values of the parameters. The results are illustrated in Figures 2.5 and 2.6 below. The input coefficients and the size distribution for the first eight (ten) investments are indicated in Figure 2.5 (Figure 2.6).

In Figure 2.6, when $\gamma_L \neq \gamma_E$, the optimal path is no longer linear and the nonproportional price development manifests itself in the form of the distribution of input coefficients.[13]

The figures show that in the dynamic case an optimal structure for an industry is a dispersed structure with units of different size and different

[12] With a more flexible production function, factor bias is also due to changes in the elasticity of substitution of the production function and non-neutral technial progress.

[13] Here we might also have indicated the size of the units. The size, however, is similar to the distribution in Figure 2.5 and has been omitted from Figure 2.6 for the sake of simplicity.

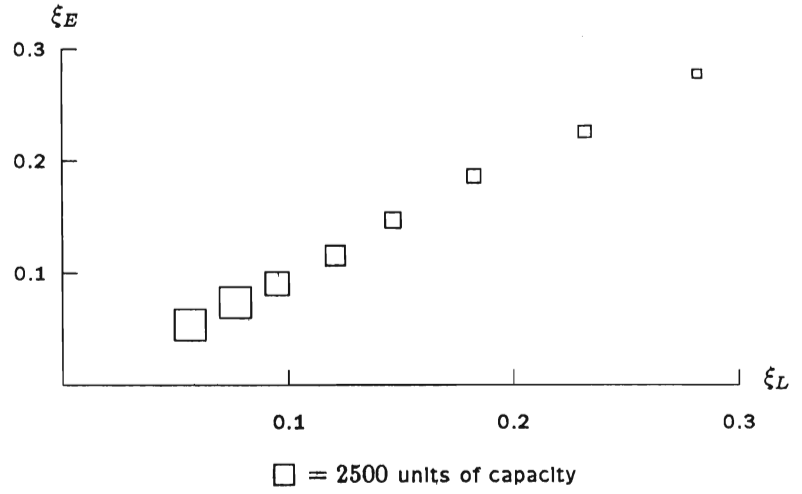


Figure 2.5: The distribution of plants with respect to input coefficients and size ($\gamma_L = \gamma_E$, $g = 0.10$).

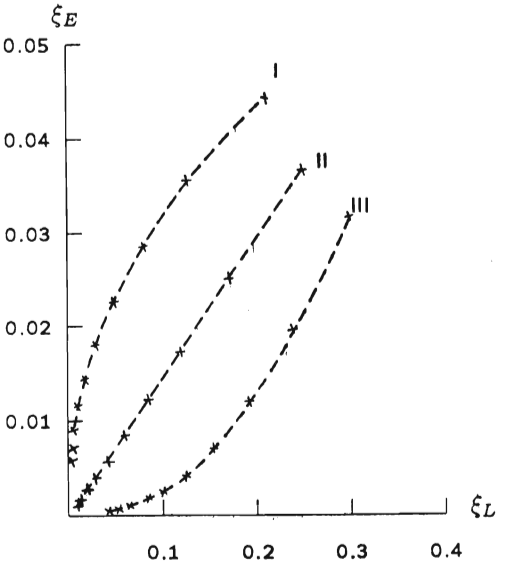


Figure 2.6: The distribution of plants with respect to input coefficients when γ_L and γ_E vary. (Path I: $\gamma_L = 0.05$, $\gamma_E = 0$. Path II: $\gamma_L = \gamma_E = 0.03$. Path III: $\gamma_L = 0$, $\gamma_E = 0.05$).

input coefficients. At any one point in time, therefore, the distribution consists of a number of points situated inside a limited interval on a path. The distribution then moves along the path so that, at a later date, the structure embraces a limited interval situated on another part of the path. In the dynamic case we can define an optimal structure as a snapshot picture of an optimal development. From a static point of view a dispersed structure may seem non-optimal, but is nevertheless part of an optimal dynamic development.

2.5 Optimal capacity expansion and the size distribution of micro units

Introduction

In the field of industrial organisation industrial structure has traditionally referred to the plant structure or firm structure of different manufacturing industries. Within this context different concentration ratios and the shape of the whole size distribution are important characteristics of industrial structure. The empirically observed size distributions exhibit a remarkable regularity — they are all highly skewed and can be fitted closely to the Pareto distribution or similar skew distributions, for example, log normal.

These common characteristics have led to speculations about the mechanism by which such distributions are generated and a number of possibilities have been explored. A somewhat disappointing conclusion is that classical production and cost theory is unable to explain the shape of the observed distribution, while a simple stochastic growth model without optimising behaviour often is successful in predicting the actual size distributions. It seems, therefore, natural to look more closely at the size distributions of micro units generated by the capacity expansion model presented in Section 2.4. Before doing so a short background and review of the topic is presented.

Background

The size distribution of firms and establishments (plants) is almost always highly skewed. It can be described graphically by the Lorenz curve. This shows the share of total business activity controlled by any given share of firms. If the curve is a straight line, all the firms are of equal size and the industry may be said to be completely unconcentrated. In general the

largest x percent of firms will control more than x percent of the activity. The Gini coefficient (the area between the diagonal and the actual curve divided by the area of the total triangle beneath the diagonal) is a numerical measure of such concentration.

Alternatively one can try to fit any distribution function to the empirical data. Whether sales, assets, number of employees, value added or profits are used as the size measure, the observed distributions always belong to the class of highly skewed distributions such as Pareto, log normal, exponential, Yule, etc. "This is true of the data for individual industries and for all industries taken together. It holds for sizes of plants as well as for firms."[14]

Attempts at economic explanations of the observed facts about concentration of industry have almost always assumed that the basic causal mechanism is the shape of the long-run average, U-shaped, cost curve (Simon and Bonini [1958], p. 607). As the scale corresponding to minimum costs need not be the same for different firms even in the same industry, firms can have the same minimum costs but varying outputs. The cost curve yields no prediction about the distribution of firms' sizes and no explanation as to why the observed distributions approximate the Pareto, log normal and other skew distributions. In the case of constant returns to scale the size distribution is undetermined. In the static analysis of economies of scale one big firm exists in long-run equilibrium.

An entirely different suggestion, for the explanation of firm size distribution is developed by Lucas [1978]. His model is based on the assumption that there is a distribution of the human ability to manage assets effectively, and consequently, a distribution in the assets that are entrusted (by the market mechanism) to each manager. The crucial point here is that the observed distribution of firm sizes is determined by the unobserved distribution of managerial ability in the population.

Since classical theory provides virtually no basis for an empirical explanation of the size distribution of firms and establishments, the search for an explanation has been directed towards stochastic processes. It is well-known that skew distributions (Pareto, Yule and log normal) can be generated by simple stochastic processes in which the so-called *Law of Proportionate Effect*, Gibrat's Law,[15] is incorporated, i.e., where current size has no effect on the expected growth of a firm. Stated more formally, in this case the *Law of Proportionate Effect* implies that the distribution of

[14] See Simon and Bonini [1958], p. 611.

[15] Originally printed in Gibrat [1931]; reprinted in Gibrat [1957].

percentage changes in size of firms in a given size class is the same for all size classes over a given time period. One can then show that incorporating the *Law of Proportionate Effect* in the transition matrix of a stochastic process results in a steady state distribution much like the skew distributions so often observed for firms and plants.

The main argument in favour of a stochastic explanation may be expressed by the following quotation:

> Since the observed distributions are radically different from those we would expect from explanations based on static cost curves, and since there appear to be no existing models other than the stochastic ones that make predictions of the shapes of the distributions, common sense will perhaps consent to what theory does not forbid accepting the stochastic models as substantially sound. (Ijiri and Simon [1964], p. 78.)

The stochastic explanation of the size distribution of firms and plants has also been prevailing. The assumption of the underlying production function has often been that of constant returns to scale. In an article on the growth of industrial concentration, Prais [1974, p. 275] also states that "the tendency towards increasing concentration is not dependent on increasing returns to scale, but is consistent in a certain sense with constant returns and even with mildly decreasing returns." The assumption of constant returns to scale is consistent with the fact that there is little or no correlation between firm sizes and profit rates, less so with empirical production studies (usually based on data for establishments or plants), which in general exhibit increasing returns to scale.[16].

In an interesting article about the relevance of classical production theory Simon [1979, p. 479] also concludes that simple stochastic growth theory does a good job of predicting the actual size distributions and that attempts which have been made to account for the observed skew distributions in terms of classical theory either fall short of the mark or require ad hoc assumptions that are not especially plausible.

Thus, according to the general opinion, "Economic theory has little to say about the distribution of firms' sizes".[17] However, in Section 1.4 we pointed out that Marx's theory of production stands out as a typical vintage theory. As a further indication, let us also quote two remarkable and most interesting passages from Marx concerning the size distribution

[16] See Singh and Whittington [1975].

[17] See Simon and Bonini [1958].

of production units:

> Under competition, the increasing minimum of capital required with the increase in productivity for the successful operation of an independent industrial establishment, assumes the following aspect: As soon as the new, more expensive equipment has become universally established, smaller capitals are henceforth excluded from this industry. Smaller capitals can carry on independently in the various spheres of industry only in the infancy of mechanical inventions. (Capital III:15, pp. 262–63.)

and:

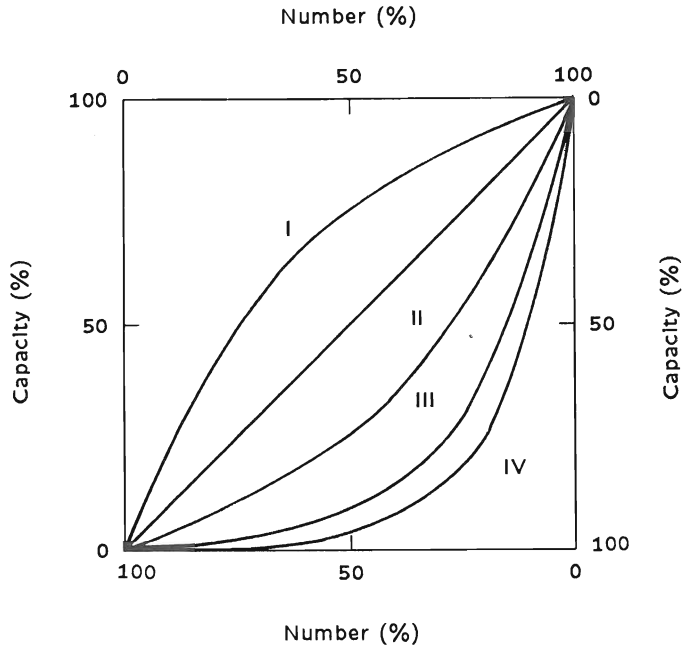
> for in each business there exists, commensurate with the development of its production, a normal minimum of invested capital essential to maintain its capacity to compete. This normal minimum grows steadily with the advance of capitalist production, and hence it is not fixed. There are numerous intermediate grades between the normal minimum existing at any particular time and the ever increasing normal maximum, a medium which permits of many different scales of capital investment. Within the limits of this medium reductions may take place, their lowest limit being the prevailing normal minimum. (Capital II:15, p. 262.)

These are, to our knowledge, the first comments on the size distribution of production units regarded as typically skew and the accompanying tendency towards increasing concentration. But the main point is that this skew size distribution and increasing concentration is a result of increasing returns to scale and market growth, two main features of the model presented in Section 2.4.

In this section we will analyse the size distribution of plants obtained from the capacity expansion model developed in Section 2.4. Furthermore, we will show that a skew distribution of production units emerges from the dynamic vintage model presented there.

From the expression of capacity increment $xe^{ng\tau}(e^{g\tau}-1)$ and $x = 100$, we have calculated the optimal capacity expansion for different time periods and values of the parameters. The size distributions of plants generated by the process of optimal capacity expansion are described in the Lorenz diagram in Figures 2.7 and 2.8. (The curves above the diagonal have their origin in the northwest corner.)

Somewhat surprisingly, the relative concentration as measured by the Gini-coefficient is observed to be roughly independent of γ and of the degree

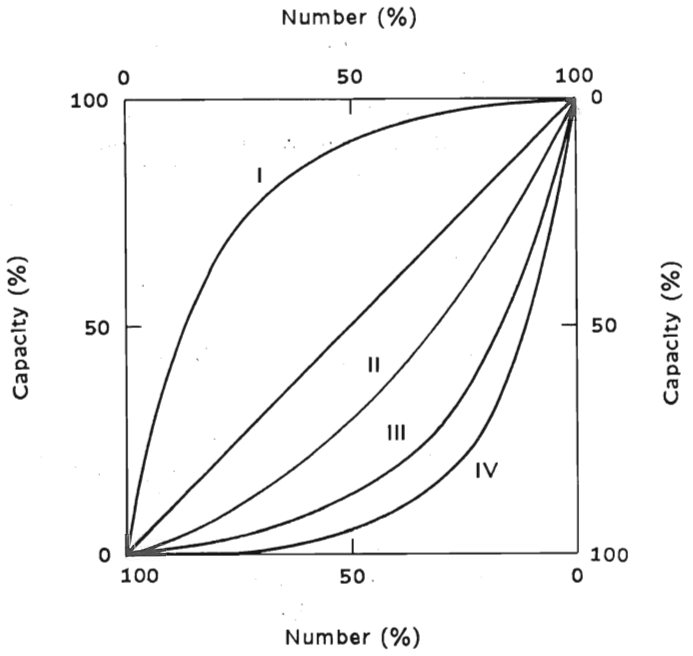


Curve I	The same curve obtained for $\gamma = -0.02, -0.05, -0.08$ and $\varepsilon = 1.25$, $g = 0.10$ or $1/\varepsilon = 0.90, 0.80, 0.70, 0.50$ and $\gamma_i = \delta = 0$, $g = r = 0.10$.
Cuve II–IV	$g = 0.03, 0.08$ and 0.12 respectively and $\gamma_i = \delta = 0$. $\varepsilon = 1.25$ and $r = 0.10$.

Figure 2.7: Relative size distribution after 50 years.

of economies of scale. Curve I in Figure 2.7 approximately holds for all the values of ε. It is the mere existence of economies of scale which yields a skew distribution, not the size. On the other hand the skewness is dependent on the time that has elapsed since the process of development started and on variations of g. As time goes by, or as g increases the skewness continuously increases.

In Figure 2.8 the curve above the diagonal shows what happens when, ceteris paribus, a finite plant life assumption of 25 years is introduced. At every investment point the capacity that is to be closed down is added to the capacity of the new plant. In this case a stationary solution results.



Curve I	Finite plant life after 50 and 70 years (coinciding curves).
Curve II–IV	Infinite plant life after 20, 50 and 70 years, respectively.

Figure 2.8: Relative size distribution for $1/\varepsilon = 0.90$, $\gamma_i = \delta = 0$, $g = r = 0.10$.

The degree of concentration is unchanged through time. Of course, we have not shown this to be an optimal policy, and complete optimisation is difficult because Theorem 2.1 no longer holds. However, this is not a serious problem in this case for the relative change in capacity for an optimal sequence of plants, compared to infinite plant life, is probably relatively small, since the capacity closed down between two investment points is very small compared to the total capacity of the new plant. The most important effect of introducing a finite plant life assumption is on the number of plants, which becomes more or less constant. This fact together with constant geometric growth sufficiently explains the stationary solution in Figure 2.8. When plant life increases, the skewness increases as well.

Distribution functions

In this section we will derive the distribution functions generated by the process of capacity expansion.

The total number of plants at time τ_n is $n+1$. At time τ_n, a plant of size x is built where x is determined by the expression for capacity expansion, $x(0,0)(e^{g\tau}-1)e^{ng\tau} = x$. For brevity $x(0,0)$ is denoted by x_0. Then one obtains

$$n = \frac{1}{g\tau} \ln\left(\frac{x}{x_0'}\right) \quad \text{where } x_0' = x_0(e^{g\tau}-1) \tag{2.75}$$

Let $F(x)$ denote the relative number of plants of size x or smaller. Thus at time τ_N

$$F(x) = \frac{n+1}{N+1} = \frac{1}{N+1} + \frac{n}{1+N} = \frac{1}{N+1} + \frac{1}{N+1} \cdot \frac{1}{g\tau} \ln\left(\frac{x}{x_0'}\right) \tag{2.76}$$

for $x_0' \leq x \leq x_0' e^{Ng\tau}$ and $F(x) = 0$ for $x < x_0'$.

If, on the other hand, $F(i)$ denotes the share of capacity due to the i largest plants one obtains an exponential distribution

$$F(i) = 1 - \frac{\sum_{n=0}^{N-i} e^{ng\tau}}{\sum_{n=0}^{N} e^{ng\tau}} = 1 - \frac{e^{(N-i)g\tau}-1}{e^{Ng\tau}-1} \approx 1 - e^{-gi\tau} \qquad (0 \leq i \leq N) \tag{2.77}$$

$$F(i) = 1 \text{ for } i = N$$

Both these distributions are typically skewed. (2.76) has certain similarities with the Pareto and also with the log normal distribution (except for small values) and (2.77) is the truncated exponential distribution.[18]

The similarity with the Pareto distribution

In the light of earlier empirical results in this field a closer look at the similarity between the derived distribution (2.76) and Pareto distribution is particularly interesting. This question is also discussed in Vining [1976b], which we draw upon here.

[18] Cf. Quandt [1966] and Ching [1973].

The distribution (2.76), here called the vintage capacity distribution, has the general form.

$$F(x) = a + b\ln(x), \qquad x_0 \leq x \leq x_1 \tag{2.78}$$

where $F(x)$ is the fraction of plants of size x or smaller, and x_0 and x_1 are the sizes of the smallest and largest plants, respectively. The constant a may be ignored when the total number of plants is large. Then, the probability density associated with (2.76) is given by the first derivative or

$$f(x) = F'(x) = b/x \qquad (x_0 \leq x \leq x_1) \tag{2.79}$$

Since $f(x)$ must integrate to one, b is given, through a simple integration, by $1/\ln(x_1/x_0)$. Thus,

$$f(x) = \frac{1}{\ln(x_1/x_0)x} \qquad (x_0 \leq x \leq x_1) \tag{2.80}$$

The cumulative distribution function and first moment associated with this density are

$$F(x) = \frac{\ln(x) - \ln(x_0)}{\ln(x_1) - \ln(x_0)} \tag{2.81}$$

$$E(x) = \frac{x_1 - x_0}{\ln(x_1) - \ln(x_0)} \tag{2.82}$$

The Pareto density, on the other hand, is given by

$$f(x) = \frac{\alpha(x_1x_0)^\alpha}{x_1^\alpha - x_0^\alpha}\frac{1}{x^{1+\alpha}} \qquad (x_0 \leq x \leq x_1) \tag{2.83}$$

The distribution function and first moment associated with the Pareto density, (2.83) are given by

$$F(x) = \frac{x_1^\alpha}{x_1^\alpha - x_0^\alpha}\left(1 - (x_0/x)^\alpha\right) \tag{2.84}$$

$$E(x) = \frac{\alpha}{\alpha - 1}\frac{x_1x_0}{x_1^\alpha - x_0^\alpha}(x_1^{\alpha-1} - x_0^{\alpha-1}) \tag{2.85}$$

Let us assume that we are given a population of plants, the largest of which is x_1 and the smallest x_0. The capacity expansion theory predicts that their sizes will be distributed approximately in accordance with (2.80) and that they will have an average size given by (2.82). On the other hand the

so-called stochastic model of firm sizes leading to the Pareto distribution predicts that they will have the distribution (2.83) and an average size given by (2.85).

If we compare (2.80) and (2.83) it turns out that the density (2.80) is a special case of (2.83) with $\alpha = 0$ and a different normalising constant.

However, the Pareto coefficient α is usually found in the range between 1.0 and 1.5 for firms. (Steindl [1965], p. 194.) Thus the density (2.80) declines with the inverse of x, while the Pareto density (2.83) declines approximately with the inverse of x squared.

The differences between the vintage capacity distribution (VC distribution for short) and the Pareto distribution will be further illustrated in the empirical section.

Some empirical results

We have succeeded in getting accurate data for a few fairly homogeneous industries in Sweden. Sweden is a small country and the number of plants in most industries is rather limited. The data for the different industries are complete, i.e., every existing plant is included. For all industries, except one, capacity data are available.

The data are described in Table 2.6. Two dominating multiplant firms are included, namely one of the two existing cement companies and the only existing sugar company. The forest-based industries are typically expanding, while sugar, cement and flour mills represent stagnating industries.

Table 2.6 shows average size and expected average size according to the VC density. The differences between the observed and the expected values are rather small. Moreover, the observed average values *exceed* the expected values.

We have also fitted the greater than cumulative for the different industries. Figures 2.9–2.17 display the empirical distribution (dots), the VC distribution (dashed line) and the Pareto distribution (solid line).

Figures 2.9–2.13 display the distribution of plant capacity for the industries in the table except sugar. The empirical distributions seem to fit the VC distribution fairly well, while they are far different from the Pareto distribution. Each of these industries consists of several independent firms, i.e., the plants are owned by several independent firms. Most forest-based firms produce a variety of products, such as sulphate pulp, sulphite pulp, paper, board, etc. In subsequent discussions about the size distribution of firm capacity, the firm is therefore a constructed unit producing a homogeneous output and including only the sulphate pulp division, the plywood

Table 2.6: Description of data for some Swedish industries and the expected value of the VC distribution.

Industry	Year	Art of data	No. of plants	Size of largest plants	Size of smallest plants	Average size	Expected value	No. of firms
1 Sulphate pulp	1973	Capacity 000 tons p.a.	31	495	25	187.9	157.4	17
2 Sulphite pulp	1973	Capacity 000 tons p.a.	33	265	5	66.7	66.5	19
3 Particle boards	1974	Capacity 000 m^3 p.a.	17	190	12	67.3	64..4	14
4 Hard board	1970	Capacity 000 tons p.a.	13	135	15	55.8	54.6	12
5 Plywood	1970	Capacity 000 m^3 p.a.	7	14	6	10.3	9.5	7
6 Cement	1968	Capacity tons p.a.	7	250	1200	674.0	606.0	2
(largest firm)	1968	Capacity tons p.a.	6	250	1200	620.0	606.0	
7 Sugar refineries	1968	Capacity tons/24hr.	6	5500	1700	3300.0	3267.0	1
8 Flour mills	1968	Production 000 tons p.a.	18	79	4	29.8	25.1	10

Source of data: Swedish Pulp and Paper, Particle Board, Wallboard and Plywood Association for industries 1–5, Ribrant (1970) for industries 6–7, and The National Price and Cartel Office (SPK) for industry 8.

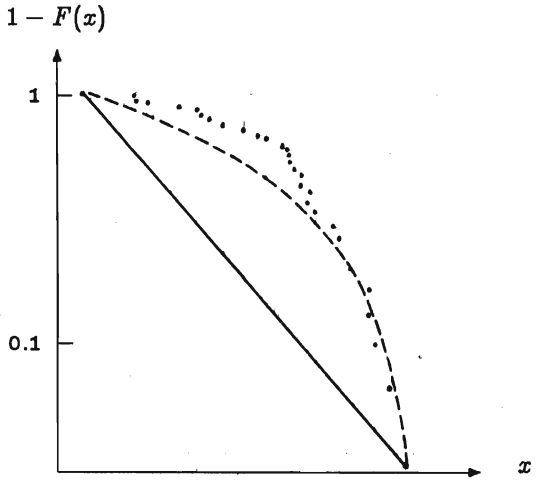


Figure 2.9: The size distribution of sulphate pulp plants.

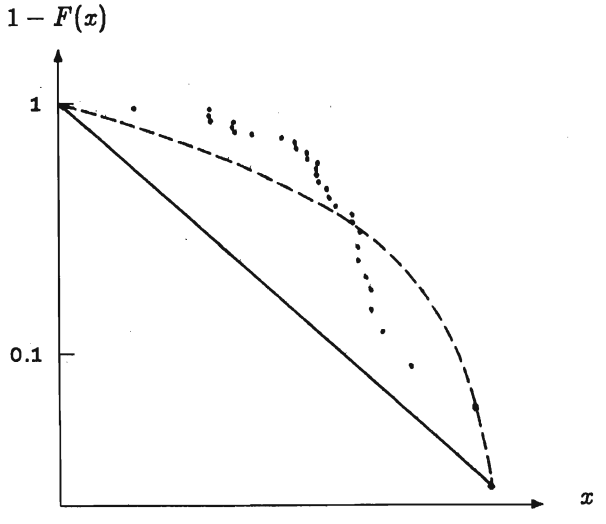


Figure 2.10: The size distribution of sulphite pulp plants.

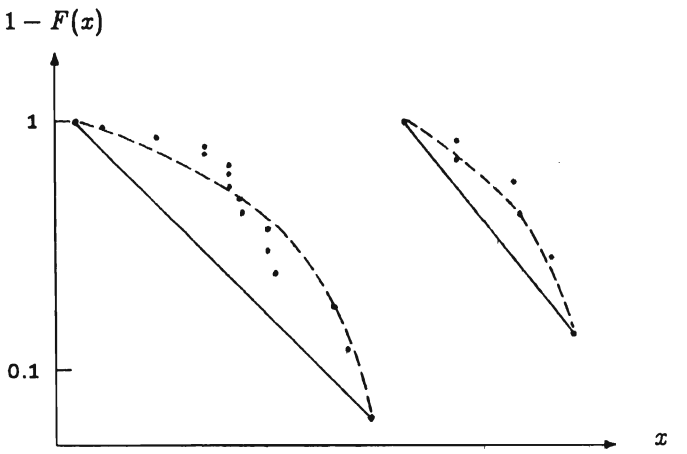


Figure 2.11: The size distribution of (a) particle board plants and (b) cement plants.

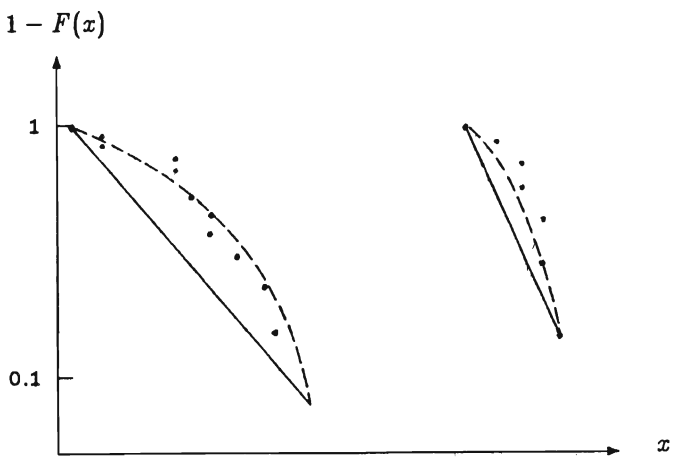


Figure 2.12: The size distribution of (a) hardboard plants and (b) plywood plants.

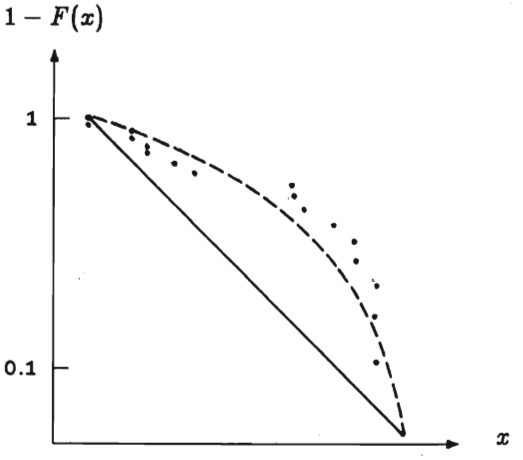


Figure 2.13: The size distribution of flour mills.

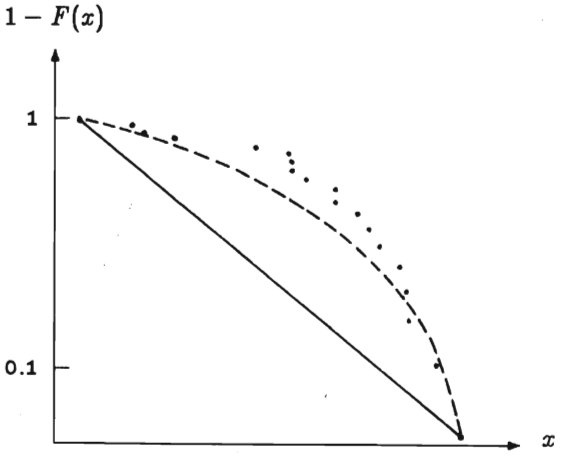


Figure 2.14: The size distribution of sulphate pulp firms.

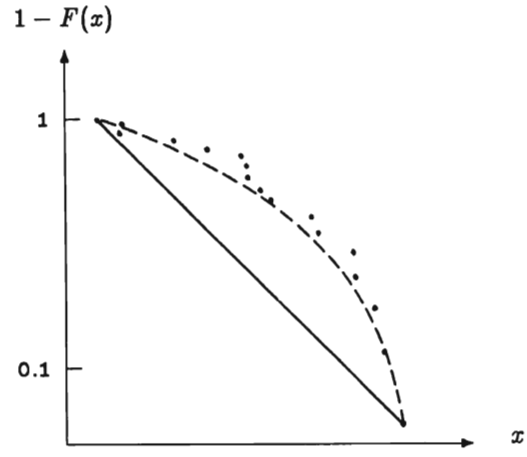


Figure 2.15: The size distribution of sulphite pulp firms.

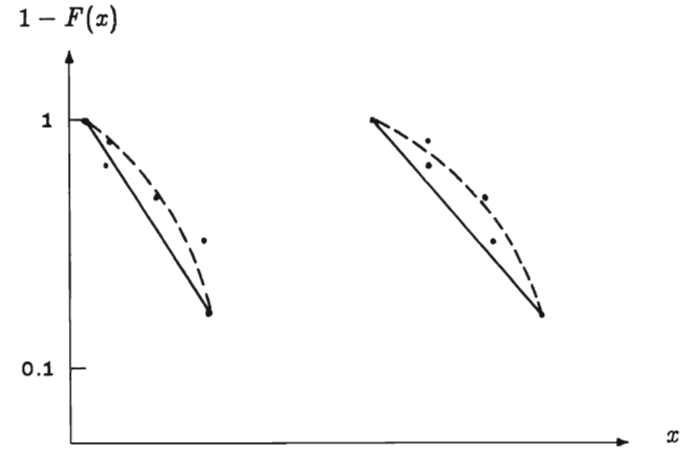


Figure 2.16: The size distribution of (a) plants within the sugar monopoly firm and (b) within the largest cement firm.

division, etc. of the real firm. The number of firms in each industry is shown in the last column of Table 2.6. In three of the industries, namely particle board, hard board and plywood, the plants belong to almost as many firms as there are plants, and the size distribution of firms does not differ very much from the size distribution of plants. In the other industries multiplant operations are more frequent. Figure 2.14 displays the size distribution of sulphate firms and Figure 2.15 the distribution of sulphite firms. In these cases the shape of the curvature is even more extreme than that of the VC distribution.

The predicted average for sulphate firms from the VC distribution is 244,000 tonnes, which differs a lot from the actual 306,000 tonnes, and for sulphite firms, 114,000 tonnes, compared with the actual 120,000 tonnes. However, the size distribution of sulphite firms fits much better to the VC distribution than that of sulphate plants. Figure 2.16 displays the distribution for two multiplant firms, the sugar company and the largest cement firm. However, these are poor examples because of the stagnation (cement) or decline (sugar) taking place in these two industries.

The realism of the model

The distributions generated by the vintage capacity expansion model are typically skew and seemingly have characteristic features similar to some empirically derived distributions. There are of course many factors which influence the size distribution (import, tranport costs, limitations of the supply of raw material, etc.) and the empirical results (grouping of data into size classes and definitions of units). Moreover, one cannot be sure that a free market generates such an optimal development.

Thus, a final but very important question concerns the nature of competition in industries with lumpy investments. The capacity structure of an entire industry is the outcome of individual firm decisions about capacity expansion. Empirical investigations of the determinants of plant sizes seem to support the view that the development of demand and aspirations for market shares are crucial for individual firms' decisions about plant sizes.[19] There are some recent efforts to develop a theory of investment competition in markets with indivisible and irreversible investments. One particular contribution which deserves attention is an article by Gilbert and Harris (1984). This article considers two different Nash-type equilibrium concepts which differ in the the type of strategies used by firms. One

[19] See, e.g., Nickell [1974] and Wohlin [1970].

model is a Cournot-Nash game, the other a preemptive competitive model. The technological aspects of the models are similar to our own. The very divergent outcomes of the two models underscore the importance of further efforts in the development of a behavioural and dynamic theory of industries with well-known empirical characteristics. [20] It should also be emphasised that it is only in the ex ante function that economies of scale are present over the entire scale. Ex post the choice is restricted to deciding the extent to which the plant is to be operated.

The behaviour of the product price influences the rate of return and pay-off period, but in vintage models above all, it also strongly influences the life span of old plants.

In spite of the simplifying assumptions made in the model, we nevertheless suppose that the model fairly well describes a typical development of what really happens when an industry changes over time. Thus it serves as a rough, first approximation which is at least as realistic as earlier models. This belief is also confirmed by several empirical investigations of the structural development of different industries in Scandinavia.[21] A more formal test of this type of model was also performed by Peck [1974] on investments in turbo generator sets. He sampled fifteen U.S. firms in the electric utilities industry for the period 1948–69 and found this model consistent with the individual firm data.

It is interesting to note that even for "socialist countries" the size distributions are skewed in a way similar to that of capitalist countries. In an article, Engwall (1972) shows that the log normal distribution fits well for enterprises in eight socialist countries, where enterprise is defined as "some hybrid of an American corporation and an American factory". This can be claimed as further support for the hypothesis that the underlying mechanism in the concentration process is basically of a technological character.

Even if the theoretical model is relevant to the size distribution of plants, it is not necessarily relevant to the distribution of firms. A more restrictive interpretation of the model might assert that it is appliable to an individual firm but not to an entire industry. Then, even if the model is relevant to the size distribution of plants within firms, it is not necessarily relevant to the distribution of firms within an industry. However, the size of a firm is determined by the size of its plants. Even if a process of mergers

[20] For a further discussion of the competitive and adjustment process under vintage assumptions in a free market, see Johansen [1972], Chapters 4 and 6, Salter [1960], Chapters IV–VII, and Section 2.2 above.

[21] See, for example, Johansen [1972], Chapter 9, and Ribrant [1970].

takes place, grouping plants into firms and firms into larger firms, the impact on the distributions or concentration measures may be very small.[22] This view is supported by the results of both Wedervang [1964] and Ijiri and Simon [1964]. Except in the upper tail, Wedervang [1964, p. 78–86] reports small differences between the distribution of establishments and firms. Ijiri and Simon show that the shape of the Pareto curve for the 500 largest firms in the USA during the past twenty years is relatively unchanged in spite of numerous mergers and acquisitions. Once an industry's structure becomes skewed through the size development of plants, it seems to require rather drastic changes in the grouping of plants into firms before the concentration measures are influenced to any considerable degree. Especially when the merger or grouping process takes place over a wide range of firm sizes the effect on the distributions seems to be of minor importance.

The conclusion drawn from our analysis is that a dynamic production and cost theory is able to explain empirical results which a static theory is unable to do. Furthermore, there exist models other than stochastic ones which generate skew size distributions. We think that our model contributes to a better understanding, and perhaps a more fundamental explanation, of what is really happening in the development of an industry over time.

2.6 Scale efficiency and the costs of decentralisation

Introduction

In this section we will utilise the model presented in Section 2.4 as a basis for a further discussion about economies of scale, scale efficiency and capacity expansion in a dynamic vintage context. Considering industrial policy in Scandinavia the trade-off between exploitation of economies of scale and increases of industrial concentration is a central question.

As mentioned in Chapter 1 industrial policy in Scandinavia has tended to stress the importance of large firms that are able to survive in international competition. In the Nordic countries, structural rationalisation policy has often promoted mergers and the construction of large production units. Such policies may be highly relevant for small open economies with small national markets insufficient to support even a single plant of

[22] Cf. also Simon and Bonini [1958, p. 612], Quandt [1966] and Scherer [1974].

optimal scale in several industries.[23] The Scandinavian attitude towards monopolies and large companies has been quite different from that of the U.S.

In contrast to the Scandinavian view and its emphasis on scale efficiency, we may refer to one of the most well-known results from standard micro theory — that competition is more efficient than monopoly. This result seems to be taken for granted in the antitrust policies of most countries. Even if most economists do not question the result, a vivid discussion about the quantitative degree of welfare losses due to monopoly has taken place in recent years. The analysis is largely based on losses of consumer surplus. The social welfare loss (the deadweight loss) arising from monopoly refers to the net reduction of consumers' surplus, i.e., the excess of the loss of consumers' surplus over the monopolist's gain in profits, the latter being regarded as a transfer of income from consumers.

The analysis of monopoly losses are usually based on average cost curves independent of market structure, despite the fact that empirical investigations have reported the existence of considerable economies of scale for most manufacturing industries. In 1968, however, the welfare trade-off between cost savings from economies of scale and the loss of consumer surplus was analysed by Williamson [1968] and formalised within a social welfare function framework. Williamson restricted his analysis to the case of a merger, which simultaneously provided cost savings and a price in excess of the competitive level. His main conclusion was that "a merger which yields nontrivial real economies must produce substantial market power and results in relatively large price increases for the net allocative effects to be negative". The purpose of this section is to look more closely at this trade-off in a vintage capacity expansion framework.

Even though the trade-off between cost savings due to economies of scale and the effects of increased market power has been analysed in the literature, the analysis is limited to traditional static price theory. By introducing putty-clay assumptions into a dynamic framework that models the capacity expansion of an industry, some dynamic efficiency aspects of monopoly and concentration in connection with economies of scale are revealed.[24]

Thus, the analysis presented here emphasises the cost level in production. In light of empirical investigations of the importance of economies

[23] For a discussion of this conflict between competitive structure and productive efficiency, see Scherer et al. [1975], Chapter 3.

[24] An elaborate treatment of this subject can be found in Hjalmarsson [1976a].

of scale in most manufacturing industries, cost aspects seem to have been rather neglected in the debate on antitrust policy and industrial concentration. We shall return to some of these problems in Section 3.6 where various aspects of static and dynamic efficiency are discussed. However, the problem of the number of firms in an industry will not be treated there.

The costs of capacity expansion

If economies of scale are present over the entire range of potential capacities of new plants, a technically optimal scale does not exist or is very large compared to demand. On the other hand, an economically optimal scale, which differs from the technically optimal scale, may nevertheless exist. In such a case, economies of scale must be treated as endogenous rather than exogenous. (The latter seems to be the rule in most analyses in which ecnomies of scale are present.) The main point now is not to achieve an optimal scale as in a comparative static analysis, but to obtain an optimal path of capacity expansion. The plant capacities generated by such an optimal process of capacity expansion are all economically optimal, even if they differ in size.

In this section we consider an industry which may consist of one or more firms, each with its own optimal process of capacity expansion and with determinate market shares constant over time. The model developed in Section 2.4 can be interpreted as a capacity expansion model for a multiplant firm producing a homogeneous product. This makes it possible to compare the costs of two different cases of capacity expansion for an industry, (1) when the capacity expansion takes place within only one, multiplant, monopoly firm and (2) when the capacity expansion takes place within an industry producing the same output, but with two or more multiplant firms.

Let us return to Equation (2.61) for the discounted stream of construction costs as well as operating costs, which can be written:

$$C(\tau) = \sum_{n=0}^{\infty} C_{\tau_n} = Hx^{1/\varepsilon} \cdot \frac{(e^{g\tau} - 1)^{1/\varepsilon}}{1 - e^{\gamma\tau}} \tag{2.86}$$

where $\gamma < 0$, $H > 0$.

$C(\tau)$ is the discounted total cost as a function of the time interval, τ, of a process of capacity expansion for an industry with an initial capacity of x and growth parameter g. H is a constant given in (2.59), and γ is a parameter given in (2.60). Let us now assume that there exists only one

firm in the industry and that it follows the rule of an optimal capacity expansion process as outlined above. Let us compare this development with that of an industry embracing two or more firms, each following the same rule of optimal capacity expansion and keeping their original market shares constant over time. The ratio of discounted costs, between the "decentralised""process of capacity expansion and the monopolistic one, is denoted by m.

Since both g and γ are assumed equal for all firms both in the multifirm case and in the monopoly firm case, τ is also assumed equal. Moreover, H is assumed equal for all firms, i.e., the same ex ante function and the same initial factor prices hold in both cases. If all firms begin their process of capacity expansion at the same moment, it is assumed they will later invest at the same points in time. This assumption is probably less realistic and tends to overestimate the value of m. The simulation of a cost-minimising development for an industry consisting of several investing firms would show that the investments should be spread over time, reducing excess capacity. In reality this may also be common in many multifirm industries, even if there seems to be a lot of exceptions, as for example the European chemical and pulp and paper industries, which seem to exhibit a very regular pattern of capacity expansion.

Let x^i be the initial capacity of firm no. i ($i = 1, \ldots, N$); $\sum_i x^i = x$, which means that the total capacity equals that of the industry with only one firm. From (2.86) one obtains a simple formula for the ratio of the discounted cost of the two capacity expansion regimes:

$$m = \frac{\sum_i x_i^{1/\varepsilon}}{x^{1/\varepsilon}} \tag{2.87}$$

From (2.86) it can then be seen that m is also the ratio between the plants' costs at every time of investment, i.e., the costs of the plants to be constructed and operated in the multifirm case, are m times those of the plant erected by the monopoly firm at the same investment point, when the capacity of the single plant belonging to the monopoly firm is equal to the aggregate capacity of the plants constructed by the multifirm industry. This also means that the average cost in the multifirm case is m times as high as the cost in the single firm case.

In Table 2.7 the value of m is calculated for different values of ε and different numbers of firms with different market shares. The number of firms is denoted by N and $x = 100$. Thus in Table 2.7 all firms are assumed to be of equal size with equal market shares.

Table 2.7: The value of m for different values of elasticity of scale. All firms have equal market shares, S_N.

N	S_N	Elasticity of scale							
		1.10	1.10	1.25	1.30	1.40	1.50	1.75	2.00
2	50	1.07	1.12	1.15	1.17	1.22	1.26	1.36	1.41
4	25	1.13	1.26	1.32	1.38	1.49	1.59	1.81	2.00
5	20	1.16	1.31	1.38	1.45	1.58	1.71	1.99	2.24
10	10	1.23	1.47	1.58	1.70	1.93	2.15	2.68	3.16
20	5	1.31	1.65	1.82	2.00	2.35	2.71	3.61	4.47
25	4	1.31	1.71	1.90	2.10	2.51	2.92	3.97	5.00
50	2	1.43	1.92	2.19	2.47	3.06	3.68	5.35	7.07
100	1	1.52	2.15	2.51	2.89	3.73	4.64	7.20	10.00

The table shows considerable differences in costs between the monopoly case and the multifirm case. These costs of decentralisation increase when the elasticity of scale increases, but decrease when the number of firms decreases. If the market shares vary between the firms, the differences in costs decrease for the same number of firms. Thus, the more unequal market shares, the less to be gained by centralised capacity expansion.

If we still assume that the values of all the parameters are the same in both cases, different optimising rules might be adopted by the firms. In markets with increasing returns to scale in the ex ante function, competition will occur over both the amount and the timing of investments. There are many possibilities with respect to the nature of firm interaction. Hence, in the multifirm case formula (2.87) probably does not hold since the time period between two investment points, τ, may differ between firms and perhaps also for the same firm over time. If the number of firms are constant, such a development will probably result in higher costs compared with the case above, in which all firms follow a cost-minimising path of capacity expansion with constant market shares. In oligopolistic markets, strategic or other considerations may result in too great an overcapacity.

Thus, we can distinguish two different aspects of efficiency here. The first one is connected with the assumption of a constant elasticity of scale

greater than 1 over the entire scale. In this case (with the assumption above) formula (2.87) shows that time does not bring anything essentially new into the analysis. The ratio m becomes independent of the time cycle τ. The same formula must also hold, ceteris paribus, when the assumption of putty-clay is removed and thus a smooth capacity adjustment in pace with demand is allowed. Inefficiency is here due to the number of firms and their market shares.

The second aspect is connected with the assumption of a putty-clay production structure and lumpy investments. When capacity expansion must take place step by step, the costs of different paths of capacity expansion become important. Inefficiency in this case is due to the lack of coordination of investment decisions, both with respect to the size of the plants and the time points of investments.[25]

Concluding remarks

Comparing our estimates of cost reduction with the trade-off-result of Williamson [1968], we find that lower average costs and monopoly welfare gains are more likely to arise in a centralised process of capacity expansion than in a decentralised one. In every case, large price increases seem to be required in the monopoly case to offset the cost reductions due to centralisation.

The analysis is limited to plant level. The effects of the firms' pricing policy of changes in market power is hard to evaluate as is also the question of increasing X-inefficiency due to monopoly. However, it is not only the competitive firm and the monopoly firm that ought to be compared, but also the whole industrial structure, the structural development and the rate of technological progress that follow a particular market type as well.

With respect to the inherent conflict in many countries between industrial policy and antitrust policy, theory is not enough. Empirical knowledge is necessary for an evaluation of the trade-off between scale efficiency and the effects of increased industrial concentration.

Appendix 2.1: Proof of Theorem 2.1

Let $\{\tau_n^*\}_{n=0}^{\infty}$ be an increasing optimal sequence of distinct points of time in constructing plants for given values of g, ε and γ.

[25] For a further discussion see Gilbert and Harris [1984].

We wish to show that $\tau_n^* = n\tau^*$, $n = 2, 3, \ldots$. The theorem can be extended to the following:

Theorem: If an optimal policy exists it is unique and has the constant cycle time property.

Proof: Examine the cost function C_{τ_n} (2.57) written in a form where the constant cycle time property is not assumed,

$$C_{\tau_n} = B\left(e^{g(\tau_{n+1}-\tau_n)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_n} \tag{A2.1}$$

Summation over all points of investment yields

$$C = B\sum_{n=0}^{\infty}\left(e^{g(\tau_{n+1}-\tau_n)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_n} \tag{A2.2}$$

Define

$$\check{\tau}_n = \tau_n - \tau_{q+1} \quad \check{\tau}_n^* = \tau_n^* - \tau_{q+1}^* \tag{A2.3}$$

for $n = q+1, q+2, \ldots$

Then the minimum of C can be written

$$\begin{aligned}\min C = \overline{C} = B\sum_{n=0}^{q-1}\left(e^{g(\tau_{n+1}^*-\tau_n^*)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_n^*} \\ + \left(e^{g(\tau_{q+1}^*-\tau_q^*)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_q^*} + e^{\gamma\tau_{q+1}^*}\sum_{n=q+1}^{\infty}\left(e^{g(\check{\tau}_{n+1}^*-\check{\tau}_n^*)} - 1\right)^{1/\varepsilon} e^{\gamma\check{\tau}_n^*}\end{aligned} \tag{A2.4}$$

Let X be the set of all vectors $x = (\tau_0, \tau_1, \ldots, \tau_{q+1}$ such that $\tau_1 < \tau_{i+1}$, for $i = 0, 1, \ldots, q$, $\tau_0 = 0$, and let Y be the set of all sequences $y = \{\check{\tau}_i\}$, $i = q+1, q+2, \ldots$ such that $\check{\tau}_{q+1} = 0$ and $\check{\tau}_i < \check{\tau}_{i+1}$ for $i \geq q+1$. Denote

$$\begin{aligned}B\Bigg(\sum_{n=0}^{q-1}\left(e^{g(\tau_{n+1}-\tau_n)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_n} + \left(e^{g(\tau_{q+1}-\tau_q)} - 1\right)^{1/\varepsilon} e^{\gamma\tau_q} \\ + e^{\gamma\tau_{q+1}}\sum_{n=q+1}^{\infty}\left(e^{g(\check{\tau}_{n+1}-\check{\tau}_n)} - 1\right)^{1/\varepsilon} e^{\check{\gamma}\tau_n}\Bigg) = W(x,y)\end{aligned} \tag{A2.5}$$

Let

$$x^* = (\tau_0^*, \tau_1^*, \ldots, \tau_{q+1}^*) \text{ and } y^* = (\tau_{q+1}^{v^*}, \tau_{q+2}^{v^*}, \ldots)$$

Then

$$\overline{C} = \min_{\substack{x\in X\\ y\in Y}} W(x,y) = W(x^*,y^*) \tag{A2.6}$$

But

$$W(x^*,y^*) \geq \min_{y\in Y} W(x*,y) \tag{A2.7}$$

On the other hand

$$\min_{y\in Y} W(x^*,y) \geq \min_{\substack{x\in X\\ y\in Y}} W(x,y) = \overline{C} \tag{A2.8}$$

That is

$$\overline{C} = \min_{y\in Y} W(x*,y) \tag{A2.9}$$

To minimise $B\sum_{n=q+1}^{\infty}\left(e^{g(\check{\tau}_{n+1}-\check{\tau}_n)}-1\right)^{1/\varepsilon}e^{\gamma\check{\tau}_n}$ over all possible sequences $\{\tau_n\}_{n=q+1}^{\infty}$ is the same problem as to minimise the expression in (A2.2) over all sequences $\{\tau_n\}_{n=0}^{\infty}$, since $\{\tau_n\}_{n=q+1}^{\infty}$ is an increasing sequence in $(0,\infty)$, with $\tau_{q+1}=0$.

Hence:

$$B\sum_{n=q+1}^{\infty}\left(e^{g(\check{\tau}^*_{n+1}-\check{\tau}^*_n)}-1\right)^{1/\varepsilon}e^{\gamma\check{\tau}^*_n} = \overline{C} \tag{A2.10}$$

Thus C can be written

$$\overline{C} = B\left[H+\left(e^{g(\check{\tau}^*_{q+1}-\check{\tau}^*_q)}-1\right)^{1/\varepsilon}e^{\gamma\check{\tau}^*_q}+e^{\gamma\check{\tau}^*_{q+1}}\frac{\overline{C}}{B}\right] \tag{A2.11}$$

where H is equal to zero if $q=0$.

Now $\check{\tau}^*_{q+1}$ by necessity minimise the expression

$$\left(e^{g(\tau-\check{\tau}^*_q)}-1\right)^{1/\varepsilon}e^{\gamma\check{\tau}^*_q}+e^{\gamma\tau}\frac{\overline{C}}{B} = \phi(\tau) \tag{A2.12}$$

where $\overline{C}$ is a constant.

Now $\phi(\tau)$ can be written

$$\phi(\tau) = e^{\gamma\check{\tau}_q}\phi(s) \tag{A2.13}$$

where

$$\phi(s) = \left(e^{gs}-1\right)^{1/\varepsilon}+e^{\gamma s}\frac{\overline{C}}{B} \tag{A2.14}$$

and

$$s = \tau - \check{\tau}^*_q \tag{A2.15}$$

From the assumption that an optimal policy exists it follows that $\phi(s)$ must have a minimum in $(0, \infty)$ which necessarily becomes unique.

The unique minimum of $\phi(s)$ is independent of q. Hence the difference $\tau^*_{q+1} - \tau^*_q$ is independent of q. This proves the constant cycle time property. Q.E.D.

Appendix 2.2: Properties of $C(\tau)$

Let us prove that $C(\tau)$ in Equation (2.61) has a unique minimum. This can be seen in the following way (γ negative by assumption):

$$C'(\tau) \underset{<}{\overset{\geq}{\equiv}} 0 \Leftrightarrow \frac{g}{\varepsilon}\frac{e^{g\tau}}{e^{g\tau}-1} \underset{<}{\overset{\geq}{\equiv}} \gamma\frac{e^{\gamma\tau}}{e^{\gamma\tau}-1} \Leftrightarrow \frac{\varepsilon}{g}(1-e^{-g\tau}) \underset{>}{\overset{\leq}{\equiv}} \frac{1}{\gamma}(1-e^{-\gamma\tau})$$

$$\Leftrightarrow h(\tau) = \frac{1}{\gamma}(1-e^{-\gamma\tau}) - \frac{\varepsilon}{g}(1-e^{-g\tau}) \underset{<}{\overset{\geq}{\equiv}} 0$$

Here $h(0) = 0$ and $h(\tau) \to \infty$ for $\tau \to \infty$.

Moreover $h'(\tau) = e^{-\gamma\tau} - \varepsilon e^{-g\tau} = e^{-g\tau}(e^{(g-\gamma)\tau} - \varepsilon)$, which is negative for $\tau < \ln\varepsilon/(g-\gamma)$ and positive for $\tau > \ln\varepsilon/(g-\gamma)$.

We see that $h'(\tau) < 0$ for $\tau < \tau_0$, $h'(\tau_0) = 0$ and $h'(\tau) > 0$ for $\tau > \tau_0$, where $\tau_0 = \ln\varepsilon/(g-\gamma)$ (which is > 0 since $\varepsilon > 1$, $g > 0$ and $\gamma < 0$).

It follows that there is a unique $\tau_1 > \tau_0$ such that $h(\tau) < 0$ for $\tau \in (0, \tau_1)$, $h(\tau) > 0$ for $\tau \in (\tau_1, \infty)$. Since $C'(\tau)$ has the same sign as $h(\tau)$, we conclude that τ_1 is a global minimum point for C.

3

The Frontier Production Function: Measurement of Productive Efficiency and Technical Change

3.1 Introduction

In Section 2.2 we introduced the concept of the ex ante production function in the vintage model. The theoretical notion behind the ex ante function is that it should show the most efficient means of transforming inputs into outputs. The blueprint technology in Grosse [1953], the best-practice technology in Salter [1960] and the ex ante function in Johansen [1972] all could be considered as fulfilling this notion. The concept itself does not suggest a unique interpretation, but is rather vague and relative.[1]

In pursuing this notion empirically, an important distinction is made between ex ante technology observed as the utilised best-practice technology in plants in operation and ex ante technology in the sense of engineering know-how not yet demonstrated in practice. Ex ante functions based on observed performance are usually called frontier production functions, while those based on engineering knowledge are called engineering production functions. The ex ante and frontier concepts are often regarded as synonymous. We may distinguish between ex ante functions according to the following criteria:

(i) current best-practice technology

(ii) blueprint technology

(iii) technology obtained through further research and development, i.e., R&D.

Concerning (iii), there is the problem of unexplored areas of technology. R&D efforts will most likely fill in the knowledge in limited parts of such

[1] For a further discussion, see Salter [1960], pp. 13–16, and Johansen [1972], pp. 6–9.

areas, but how far the ex ante concept should be taken in this direction is still an open question.

In this study we shall consider frontier functions which are based on observed performances. Engineering approaches to ex ante functions are, of course, highly relevant, especially for production units at a disaggregated level. However, they are usually outside the economist's area of competence, and necessitate considerable efforts compared with estimations on observed data.[2]

The frontier production function is used to answer the following questions:

(i) How much output can be expected when new production capacity is introduced in the industry?

(ii) How do the units within the industry perform using a frontier function as a basis of comparison?

3.2 Definition of the frontier production function

In this study we are solely concerned with industries producing a single homogeneous good. Consider an industry with N firms or plants, all producing a single homogeneous output x from a vector of inputs v, v consisting of current inputs and capital. The production possibilities are described by a set of production functions

$$x^j = f^j(v^j), \quad x^j \in R_+, \quad v^j \in V, \qquad V \subset R_+^n \qquad j = 1, \ldots, n \quad (3.1)$$

These production functions represent the blueprint technologies from which the choice of technique was selected at various construction dates. The best-practice or frontier production function (in a factor space $V \subset R_+^n$), for an entire industry consisting of a given set of N firms or plants with production functions according to (3.1), is defined by

$$F(v) = \max_j f^j(v^j), \quad v^j \in V \qquad j = 1, \ldots, N \quad (3.2)$$

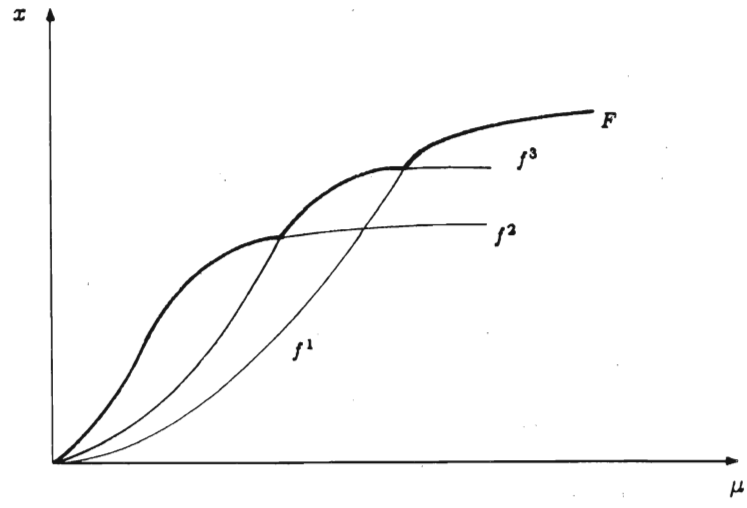
The frontier production function is made up of those parts of the firms' production functions that yield maximum output for a given set of inputs,

[2] See Eide [1979] for the derivation of an ex ante function for oil tankers based on engineering simulation design models.

relative to the set of production functions applying to the industry.[3] A special case is when one function is identical with the frontier function. The frontier function is continuous if the firm functions are continuous, but not necessarily differentiable at every point.

Since no restrictions have been imposed on the available amount of capital, this frontier function concept is not identical to the relationship obtained when maximising total industry output subject to a given total amount of inputs, with the micro units' set of production functions subject to the actual capacity constraints. This latter concept will be termed the short-run industry function and developed extensively in Chapter 5.

Figure 3.1 illustrates the concept in the case of three firms. By cutting the production functions with a vertical plane through the origin (i.e., μ indicates a factor ray), Figure 3.1 shows that, depending on the scale of operation, some part of each of the production functions belongs to the frontier production function.



The frontier production function F as an envelope of three individual ex ante functions, f^j.

Figure 3.1: The frontier production function.

[3] See, for example, Aigner and Chu [1968].

3.3 The measurement of efficiency

The notion of efficiency

The concept of efficiency is, in a broad sense, used to characterise the utilisation of resources. In other words, efficiency is a statement about the performance of processes in transforming a set of inputs into a set of outputs. Efficiency is a relative concept, that is, the performance of an economic unit must be compared with a standard. Establishing a standard involves value judgments with respect to the various objectives pursued by economic units.

The choice of specific efficiency measures depends on the purpose of the measurements. Efficiency measures are usually applied at the following three levels of aggregation:

The macro level

Efficiency measures are used at an aggregate level to indicate allocative efficiency, i.e., the economic performance of an observed allocation of resources to different sectors is compared with the result of some ideal allocation. A usual exercise is to measure the loss due to monopoly. The ideal allocation is usually required to be Pareto-optimal, given the existing income distribution. Another standard of reference is an allocation that maximises some welfare function.

The industry level

The purpose here is to measure the relative performances of the firms within an industry, and thereby, to give a picture of the structure of the industry. The notion of a best-practice firm or a frontier production function serves as a measuring rod for performance. Efficiency measures at this level show the potential for an increase in industry output by employing resources in firms using best-practice technology.

The micro level

Efficiency at the level of a single firm concentrates on the utilisation of resources within the firm. The measures at the industry level are based on given sets of production possibilities for each of the firms. The problems at the microeconomic level are the managerial and engineering problems

of reaching the maximum output for a given set of inputs. A best-practice technology is also the reference at the microeconomic level. Obviously, the particular objectives of a firm must be specified when characterising its efficiency, i.e., a firm can be perfectly efficient with respect to its own objectives, but inefficient with respect to other objectives that the investigator decides is superior.

The efficiency frontier

Efficiency measures are often based on unit requirements of inputs, i.e., the production functions are transformed from the factor space into a space of input coefficients

$$\begin{aligned} \xi &= (\xi_1, \ldots, \xi_n), \quad \xi_i = v_i/x \qquad i = 1, \ldots, n \\ x^j &= f^j\left(\frac{v^j}{x^j} x^j\right) = f^j(\xi^j x^j) \qquad j = 1, \ldots, N \end{aligned} \tag{3.3}$$

This transformation forms a set of feasible input coefficients bounded towards the origin and the coordinate axes of the factor space under certain restrictions on the forms of the establishment production functions. A sufficient restriction is that the functions conform to the *regular ultra passum law.*[4]

For homogeneous functions the set of input coefficients collapses into a single curve in the case of constant returns to scale. The set of input coefficients is not bounded for functions homogeneous of a degree not equal to 1.

Considering only a single production function such as (3.1), the elasticity of scale is defined by the *passus equation*[5]:

$$\varepsilon = \frac{\sum_{i=1}^{n} (\partial f/\partial v_i) v_i}{f} = \Phi(v_1, \ldots, v_n) \tag{3.4}$$

If $f(\cdot)$ is continuously differentiable, then the scale elasticity is a continuous function of the inputs. It follows directly that ε is a directional elasticity.

For a proportional factor variation

$$v_i = \mu v_i^0 \qquad i = 1, \ldots, n \tag{3.5}$$

[4] Defined by Frisch [1965], the law implies that the elasticity of scale is decreasing along arbitrarily rising curves in the input space from values greater than 1 to values smaller than 1. For further analysis, see Appendix 3.1.

[5] See Frisch [1965], Ch. 8, and Danö [1966], Ch. IV.

where v_0^i is a given point and μ is a positive scalar. Inserting (3.5) in (3.1) provides an equivalent definition of elasticity of scale

$$\varepsilon = \frac{df/f}{d\mu/\mu} = \varepsilon(\mu) \tag{3.6}$$

The technically optimal scale is defined as the locus of all points where average productivities reach their maximum value on each ray through the origin. We now look for extreme values of the average factor productivities, x/v_i, along the factor ray

$$\frac{\partial(f/v_i)}{\partial\mu} = \frac{v_i(df/d\mu) - f(dv_i/d\mu)}{v_i^2} = \frac{v_i^0 f(\varepsilon - 1)}{(\mu v_i^0)^2} = 0 \tag{3.7}$$

The average productivities under proportional factor variation will have extreme values when the elasticity of scale is equal to 1. A sufficient second-order condition for a maximum is

$$\begin{aligned} d^2(f/v_i)/d\mu^2 = \big[v_i^2\big(v_i^0(f\,d\varepsilon/d\mu + \varepsilon df/d\mu - df/d\mu)\big) \\ - v_i^0 f(\varepsilon - 1)2v_i dv_i/d\mu\big]/v_i^4 = (v_i^0 f/v_i^2)d\varepsilon/d\mu < 0 \\ i = 1, 2, \ldots, n \end{aligned} \tag{3.8}$$

The assumption of a regular ultra-passum law ensures that the average productivities have uniquely determined maximum values for $\varepsilon = 1$ when moving along any factor ray, since $d\varepsilon/d\mu < 0$ by definition.

The geometric locus of such points where $\varepsilon = 1$ for different factor rays is defined as the surface of technically optimal scale. From (3.4) the equation for the optimal scale results in

$$\Phi(v_1, v_2, \ldots, v_n) = 1 \tag{3.9}$$

The optimal scale is an $n-1$ dimensional surface in the factor space.

Since the input coefficients are the inverse of average productivities, the transformed optimal scale curve must be the boundary towards the origin and axes of the set of feasible input coefficients. The optimal scale surface is transformed into the input-coefficient space by inserting $v_i = \xi_i x$ into (3.9). Using the relationship implicit in (3.3) to express x as a function of the ξ's and substituting, we can then derive a relationship between the unit requirements on the optimal scale function in the input-coefficients space which, analogous to (3.9), is written as

$$\Psi(\xi_1, \xi_2, \ldots, \xi_n) = 1 \tag{3.10}$$

The optimal scale surface transformed to the input-coefficient space is called the efficiency frontier.[6]

Let us now return to the set of production units comprising an industry. Assuming functional forms resulting in input-coefficient sets bounded towards the origin and the coordinate axes of the factor space, the efficiency frontier E for an industry consisting of N firms with production functions described by (3.3) is defined by

$$E = \left\{ \xi = (\xi_1, \ldots, \xi_n) | \xi_k = \min_j \min_\mu \frac{\mu v_k^0}{f^j(\mu v^0)} \right.$$
$$\left. k = 1, \ldots, n, \quad j = 1, \ldots, N, \quad \mu v^0 \in V, \ \mu \in (0, \infty) \right\} \tag{3.11}$$

where μv^0 denotes a factor ray.

The efficiency frontier is made up of all points where the input coefficients $(\xi_1, \ldots, \xi_n)$ reach their minimum values along rays from the origin through μv^0. Under our regularity assumptions all such efficiency frontier points are boundary points of the feasible production set.

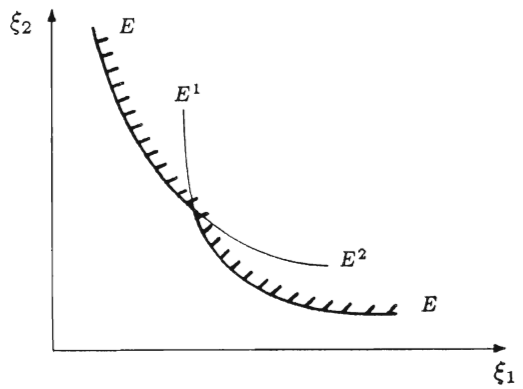


Figure 3.2: The efficiency frontier as an envelope of two transformed optimal scale curves.

An illustration is provided in Figure 3.2 for the case of two inputs. For the transformation of one firm function, say No. 1, the efficiency frontier, E^1, corresponding to this function represents the optimal scale of the function,

[6] In Johansen [1972], p. 21, it is also called the technique relation.

i.e., the scale elasticity is equal to 1 on the frontier. The input coefficients reach their minimum values subject to proportionate variation of the inputs when the elasticity of scale is equal to 1.[7] The efficiency frontier EE is identical to the curve corresponding to the optimal scale of the frontier production function.

3.4 Generalised Farrell measures of efficiency

Farrell measures

In the seminal paper by Farrell [1957] three types of efficiency measures were introduced: technical efficiency, price or allocative efficiency, and overall efficiency.

Farrell assumed that one single production function with constant returns to scale represented the entire frontier production function. In such a case the transformed isoquants collapse into one single curve in the input-coefficient space.

Following Farrell, *technical efficiency* is measured by comparing observed input-coefficient points for a firm with the input coefficients on the efficiency frontier for the same factor proportions. The two-input case is shown in Figure 3.3.

Technical efficiency of a firm with observed input coefficients represented by D is measured by the ratio OA/OD. This measure shows the relative reduction in input requirements by producing the observed output with frontier production technology and the same factor proportions.

When measuring *allocative efficiency*, i.e., when passing judgment about the combination of inputs, the standard of reference must be based on some objective function, either the firm's own or the investigator's. Assuming that all the firms face the same factor prices and that the objective is to minimise costs, a measure of allocative efficiency, or *price efficiency*, is based on comparing observed average cost with the average cost represented by the unit-cost line through E and C in Figure 3.3, which in turn is the result of using cost-minimising factor proportions. In the case of functions that are homogeneous of degree 1, price efficiency for firms with observed factor proportions represented by the ray OD in Figure 3.3 is measured by the ratio of average cost representing cost minimisation and

[7] See Førsund [1971].

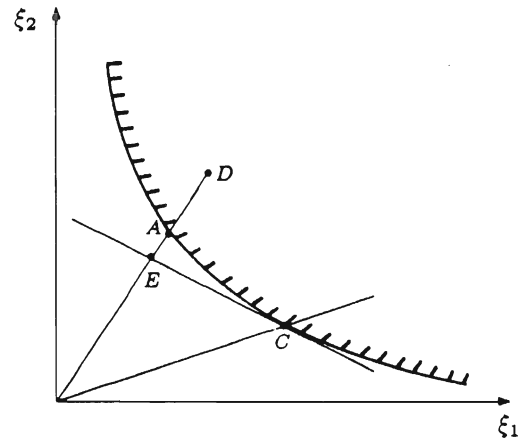


Figure 3.3: An illustration of Farrell's efficiency measures.

the average cost of the technically efficient firm for that factor proportion, i.e., OE/OA.

It should be noted that the cost-minimising proportions are independent of the scale of production only in the case of homothetic production functions. Farrell's measure of price efficiency is therefore of limited interest.

Farrell combines these technical and price efficiency measures by taking the product of the two measures, i.e., referring to Figure 3.3, overall efficiency is measured by the ratio $(OA/OD) \cdot (OE/OA) = OE/OD$. The inherent weakness of the price efficiency measure then also applies to the overall efficiency measure.

Generalised Farrell measures

In this section the Farrell measures are generalised to non-homogeneous production functions, still assuming that a single production function represents the entire frontier production function.

The measures are radial, i.e., the distance between an observed unit and the reference path is measured along a factor ray. This can generally be justified by the splitting of total efficiency into two components, one showing potential cost reductions due to a proportional movement along a factor ray (technical efficiency and scale efficiency) and another showing

the potential cost reduction due to movement along an isoquant (price efficiency). In this study we are not concerned with price efficiency.

It is not obvious that the analysis of efficiency should be limited to ray measures. In Färe and Lovell [1978] an approach based on minimising the value of various distance measures from an observation to an efficiency frontier was developed.[8] This approach was extended to the case of multiple outputs in Färe et al. [1983].[9] A strong argument in favour of applying the radial measures is that these measures have a straightforward economic interpretation, and in addition, are continuous.[10]

Assuming that an efficiency frontier exists, the frontier production function and the efficiency frontier are illustrated in Figure 3.4, and for the two factor case in Figure 3.5. These figures will also be used to illustrate the different measures of efficiency.

Figure 3.4 relates to a unit observed to have inputs v^0 and output x^0 at D'. A section of the production function is represented by the curve $x = f(\mu v^0)$. Output per unit of input is maximised when a ray from the origin is tangential to $f(\mu v^0)$ as at A', where output is $\hat{x}$ and the scale elasticity ε is unity. This is the technically optimal scale. B' and C' are points on $f(\mu v^0)$ respectively corresponding to a unit producing the observed output x^0 with minimum inputs $\mu_1 v^0$ and to a unit producing maximum output x^* with actual inputs v^0. Minimum and maximum refer to the frontier technology.

In Figure 3.5 optimal scale of the production function is transformed to the input-coefficient space. Point A, corresponding to A' in Figure 3.4, lies on the efficiency frontier. B and C are the transformed points of B' and C' in the production surface of Figure 3.4, corresponding to output levels x^0 and x^*, respectively. D is the observed point $(v_1^0/x^0, v_2^0/x^0)$ corresponding to D'. The slope of the ray OD is v_2^0/v_1^0.

Technical efficiency

Two different measures of technical efficiency denoted by E_1 and E_2 may be defined when allowing for production functions homogeneous of a degree different from 1. An illustration of these measures is provided in Figures 3.4 and 3.5.

---

[8] Cf. further discussion in Kopp [1981a, 1981b], and Färe and Lovell [1981].

[9] See also Färe et al. [1985] and Charnes et al. [1978] and [1981].

[10] See Russell [1985a] and [1985b].

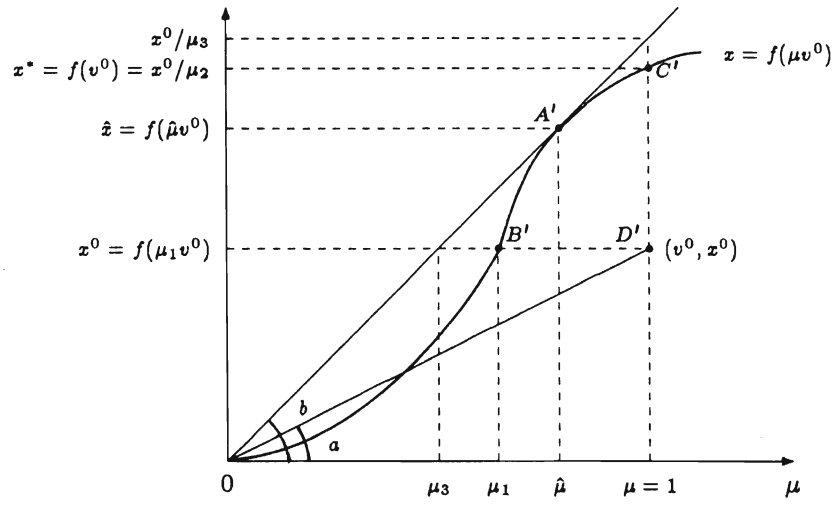


Figure 3.4: A section of the frontier production. Function $x = f(v)$ along the ray $v_i = \mu v_i^0$.

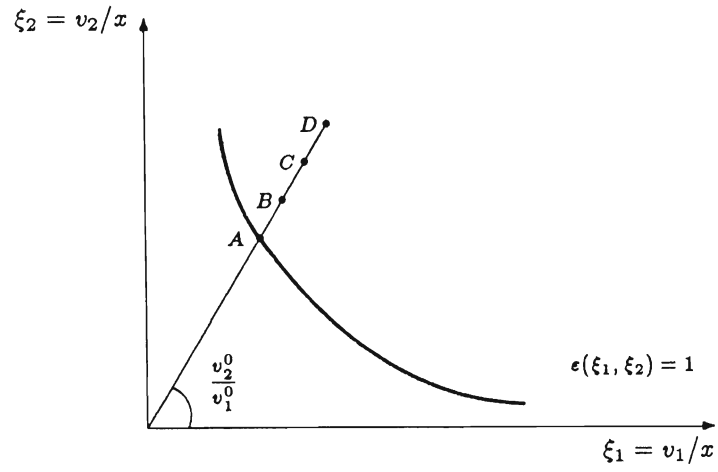


Figure 3.5: The efficiency frontier.

The *input saving measure*, E_1, is obtained by comparing an observed point of input requirements and output (v^0, x^0) with the input requirements on the frontier production function corresponding to the observed output. Looking at Figure 3.4, the observed point at D' and the point on the frontier production function directly yields

$$E_1 = \mu_1 \tag{3.12}$$

where μ_1 is found by solving for μ_1 in $x^0 = f(\mu_1 v^0)$. This measure shows the ratio between the amount of inputs required to produce the observed output with the frontier function technology and the observed amount of inputs. In the input-coefficient space this would mean comparing an observed input-coefficient point with the point on the transformed isoquant of the frontier function corresponding to the observed output and the observed factor proportions. By definition this transformed isoquant must lie closer to the origin. The measure will then show the relative reduction in the amount of inputs needed to produce the observed output, using the frontier function technology and the observed factor proportions. In Figure 3.5,

$$E_1 = OB/OD \tag{3.13}$$

The *output increasing measure*, E_2, is obtained by comparing an observed point D' of input requirements and output (v^0, x^0), with the output obtained on the frontier production function for the same amount of inputs at point C'. Referring to Figure 3.4,

$$E_2 = \frac{x^0}{x^*} = \frac{f(\mu_1 v^0)}{f(v^0)} \tag{3.14}$$

This measure shows the ratio between the observed output and the potential output obtained by employing the observed amount of inputs in the frontier function.

In the input-requirement space this means comparing an observed point with the point on the transformed isoquant of the frontier production function corresponding to the output obtained by employing the observed amount of inputs in the frontier function. In Figure 3.5

$$E_2 = OC/OD \tag{3.15}$$

These two measures, E_1 and E_2, will generally not coincide except in the case of linear homogeneity. However, there is an interesting relationship between E_1 and E_2 and the elasticity of scale (or the *passus coefficient* as

Frisch called it). In Frisch [1965, p. 73], there is an identity called the *second form of beam variation equation*, which shows that under proportional variation of inputs the proportionality factor μ can be multiplicatively separated

$$f(\mu v_1^0, \mu v_2^0, \ldots, v_n^0) \equiv f(\mu v^0) = f(v^0) \cdot \mu^{\bar{\varepsilon}(\mu)} \tag{3.16}$$

where $\bar{\varepsilon} = \int_\mu^1 \frac{\varepsilon(\tau)}{\tau}\, d\tau / \int_\mu^1 \frac{1}{\tau}\, d\tau$, which is a weighted average of the elasticity of scale in the interval between x^0 and x^* in Figure 3.4.

Rearranging (3.16) yields

$$\frac{f(\mu v^0)}{f(v^0)} = \mu^{\bar{\varepsilon}} \tag{3.17}$$

or

$$\bar{\varepsilon} = \frac{\ln\left(\frac{f(\mu v^0)}{f(v^0)}\right)}{\ln \mu} \tag{3.18}$$

Substituting in E_1 and E_2 we obtain

$$\bar{\varepsilon}(\mu) = \frac{\ln E_2}{\ln E_1} \tag{3.19}$$

or

$$E_2 = E_1^{\bar{\varepsilon}(\mu)} \tag{3.20}$$

Thus $E_1 \underset{<}{\overset{\geq}{=}} E_2$ for $\bar{\varepsilon}(\mu) \underset{<}{\overset{\geq}{=}} 1$. As stated above, the two measures coincide when f is homogeneous of degree 1.

The ranking of units according to the two measures of technical efficiency coincides if the elasticity of scale is constant or does not pass through the value 1 in the sample. Since we have chosen E_1 and E_2 to be numbers with values between 0 and 1, E_1 is greater (smaller) than E_2 when the average of the elasticity of scale is greater (smaller) than 1. Thus in Figure 3.5 we have arbitrarily chosen $E_1 < E_2$.

In empirical studies the choice between the two measures should be determined by the objective. If the amount of resources is assumed to be fairly constant, e.g., a fixed total employment, then E_2 is the relevant measure; and if output is assumed to be constant, then E_1 is the relevant measure.

Scale efficiency

A measure of scale efficiency shows how close an observed firm actually is to the optimal scale. Three different measures of scale efficiency are defined

here. These measures are of course dependent on the existence of a unique efficiency frontier. (This is not the case for the technical efficiency measures.) They are of special interest in a long-run analysis of the potential possibilities for increased productivity.

The first measure of scale efficiency, E_3, shows in terms of the input-coefficient reduction the distance from an observed firm to the optimal scale on the frontier function, i.e., the ratio of an input coefficient evaluated at the technically optimal scale for the observed input ratios at A' to the corresponding observed input coefficient at D'. In Figure 3.4 the input coefficients ξ_i $(i = 1, \ldots, n)$ are constant and equal to the observed ξ_i^0 along the ray OD' and constant and equal to the coefficients obtained at optimal scale $\hat{\xi}_i$ along the ray OA'. Let a be the slope of OD' and b the slope of OA'. These slopes, equal to average productivities, may then be utilised to give the following expressions for E_3,

$$E_3 = \frac{\hat{\xi}_i}{\xi_i^0} = \frac{a}{b} = \frac{x^0}{\hat{x}/\hat{\mu}} = \mu_3 \tag{3.21}$$

where the last expression follows from the simple geometrical relationship

$$\frac{\hat{x}}{\hat{\mu}} = \frac{x^0}{\mu_3} \tag{3.22}$$

In Figure 3.5 we have

$$E_3 = OA/OD \tag{3.23}$$

The interpretation of this measure is the relative reduction in input coefficients made possible by producing at optimal scale on the frontier production function with the observed factor proportions.

E_3 is not a measure of pure scale efficiency. To obtain such a measure one has to eliminate the technical inefficiency of the observations by moving each observed unit to the surface of the frontier function. This can be done in two different ways corresponding to the two definitions of technical efficiency, i.e., by moving the units to the frontier either in the vertical or in the horizontal direction in Figure 3.4.

When moving a unit in the horizontal direction the second measure of scale efficiency, E_4, shows the distance from the transformed isoquant, corresponding to x^0, to the optimal scale. In Figure 3.5

$$E_4 = OA/OB \tag{3.24}$$

When moving a unit in the vertical direction the third measure of scale efficiency, E_5 shows the distance from the optimal scale to the transformed isoquant, corresponding to x^*. In Figure 3.5

$$E_5 = OA/OC \tag{3.25}$$

The interpretation of E_4 and E_5 is the relative reduction in input coefficients by producing at optimal scale on the frontier function with the observed factor proportions of a firm whose technical inefficiency has been eliminated in two different ways, corresponding to the definition of E_1 and E_2, respectively.

From the definition of the efficiency measures (3.13), (3.15), (3.23), (3.24) and (3.25) it follows easily (see Figure 3.5) that

$$E_4 = E_3/E_1 \tag{3.26}$$

$$E_5 = E_3/E_2 \tag{3.27}$$

Since the efficiency frontier constitutes the limit towards the origin of the feasible input coefficients, E_3 is always smaller than E_1 and E_2, except for units producing exactly at optimal scale on the frontier production function.

From (3.19), (3.26) and (3.27) we also find that

$$\bar{\varepsilon} = \frac{\ln E_3 - \ln E_5}{\ln E_3 - \ln E_4} \tag{3.28}$$

This formula shows the relationship between the scale elasticity and the three different measures of scale efficiency. Thus, all measures of scale efficiency can be expressed as a function of the average elasticity of scale.

One must remember here that the average elasticity of scale $\bar{\varepsilon}$ depends on the observation chosen, i.e., a specific $\bar{\varepsilon}$ is obtained for each observation.

Structural efficiency

In his original article Farrell also suggested a measure of technical efficiency of the whole industry, i.e., a measure of structural efficiency, by simply taking a weighted average (by output) of the technical efficiencies of the industry's constituent production units. We have extended the Farrell analysis on this point. Several other measures of structural efficiency therefore are introduced below.

According to Farrell [1957, p. 262], the purpose of a structural efficiency measure is to measure "the extent to which an industry keeps up

with the performance of its own best firms". In our context we want the structural measures to reflect the same for the industry as the individual efficiency measures show for a micro unit, i.e., potential input saving, E_1, potential increase of output, E_2, and potential reduction in input coefficients, E_3, E_4 and E_5.

The approach suggested by Farrell is to weight the individual measures by observed output levels. Thus, the first measure of structural efficiency, here denoted by S_0, is obtained by taking the average of the E_1 technical efficiency measures with outputs as weights. However, the main problem with this approach is that the result of this weighting scheme does not have a straight-forward interpretation in terms of the objectives of the structural measures, i.e., in terms of resource saving or output increasing.

Another approach (indicated by Farrell's qualifications on the weighted measure) is to construct an average firm for the industry, regard this average firm as any other observation and then compute E_1, E_2 and E_3 for this average unit. (Here we construct the average firm by taking the *arithmetic* average of each amount of inputs and outputs). These measures of structural efficiency are denoted by S_1, S_2 and S_3, where S_1 and S_2 are measures of structural technical efficiency and S_3 is a measure of structural scale efficiency.

These last three measures seem to be more satisfactory as measures of structural efficiency as specified above than the S_0 measure, since the former measures may be explicitly interpreted in terms of input saving or output augmenting for the industry. However, the reason for calculating S_0 is that it seems to be the only measure of structural efficiency that has been used in earlier studies.[11]

By adjusting the average firm to the frontier in the two different ways corresponding to the E_1 and E_2 measures, the elimination of structural technical inefficiency yields two other measures of pure structural scale efficiency corresponding to E_4 and E_5 and denoted by S_4 and S_5. It is obvious that

$$S_4 = S_3/S_1 \tag{3.29}$$

and that

$$S_5 = S_3/S_2 \tag{3.30}$$

Even in this case there exists a clear relationship between the scale properties of the production function and the efficiency measures. Since the

[11] See, e.g., Carlsson [1972].

average unit can be regarded as an arbitrary observation, the relationship between the different measures of structural efficiency and the average of the elasticity of scale is the same as the relationship between the corresponding E_i measures. Thus,

$$\bar{\varepsilon} = \frac{\ln S_2}{\ln S_1} \tag{3.31}$$

and

$$\bar{\varepsilon} = \frac{\ln S_3 - \ln S_5}{\ln S_3 - \ln S_4} \tag{3.32}$$

Because of the analogy with the E_i measures, S_3 always shows a lower value than S_1 or S_2, except in the case where the industry consists of a number of firms of optimal size employing the same best-practice technique, a situation characterising a long-run equilibrium of an industry.[12]

While the relationship between S_1 and S_2 is given by (3.31) it is difficult to analytically determine how S_0 is related to the other measures. Constructing an average unit of units with $E_1 = 1$ yields a new unit with $E_1 < 1$ if they have different factor ratios, i.e., the frontier units tend to contribute more to the S_0 measure than the S_1 measure does. The relative impact on S_0 and S_1 of units below the frontier is difficult to assess.

When the large units are on or near the frontier one may expect that S_0 is larger than S_1 due to the weighting by output shares. In the empirical results of Chapter 7, however, S_0 is always greater than S_1, even in the year when the largest unit has the lowest E_1 measure. This illustrates the impact of the structure as a whole on the differences between the measures.

3.5 Dynamic aspects of efficiency

Static versus dynamic efficiency

As shown in Section 2.3, if the underlying technological structure is characterised by ex post rigidity of factor proportions and embodied technical progress, one should be particularly careful not to attach undue normative significance to Farrell's concepts or measures. These static efficiency concepts can be rather misleading, giving a deceptive appearance of perpetual

[12] See Section 2.3 for a discussion of optimal structure and structural change of an industry and long-run equilibrium.

dissatisfaction with existing structure which has no basis from a dynamic perspective.

The actual production possibilities of an industry at any given moment of time are determined not by the latest *ex ante* function, but rather depend on the technology and capacities of all the existing production units.

Referring to Figure 2.3, we may realize that the production units are concentrated at one point in the input-coefficient space only under very special circumstances. In short, we may say that these are circumstances characterising a steady state (constant *ex ante* function, constant factor prices and no wear and tear that makes production with old equipment more input-consuming than production with new equipment). In this rare case static and dynamic efficiency do coincide.

In a diagram of input-coefficients, a more usual or representative picture of an industry would reveal a dispersed structure with units of different size and input coefficients.[13] The existing structure is under constant development due to the scrapping of old units and the choosing of new technology that occur under new investment.

When specifying an objective function for industrial policy — for instance, cost minimisation — an optimal structural development can be derived as shown in Chapter 2. Inefficiency can then be measured on the basis of such an optimal development. In this case we define an optimal, or efficient, structure as a snapshot phase of an optimal development. A dispersed structure which may be inefficient from a static point of view, nevertheless, may be part of an optimal dynamic development. From a policy point of view the problem is not to bring the existing structure closer to the best practice structure, but to optimise a process that is going on all the time.

A diagram of input coefficients for an industry is used to describe this process. Figure 3.6 shows two different efficiency frontiers that have existed in the past, E_{t-1} and E_{t-2}, the actual E_t and two estimated future efficiency frontiers, E_{t+1} and E_{t+2}, the latter representing a technological forecast. In the case of perfect efficiency all the production units in the industry should be situated on the optimal path at the intersections between the path and the efficiency frontiers E_t, E_{t-1} and E_{t-2}.

The actual existing units are indicated by open circles. If all firms consist of only one production unit, the figure also shows the firm distribution. On the other hand, if a firm consists of several production units, the input coefficients of the firm are derived as weighted averages of the

[13] See Johansen [1972] and Salter [1960].

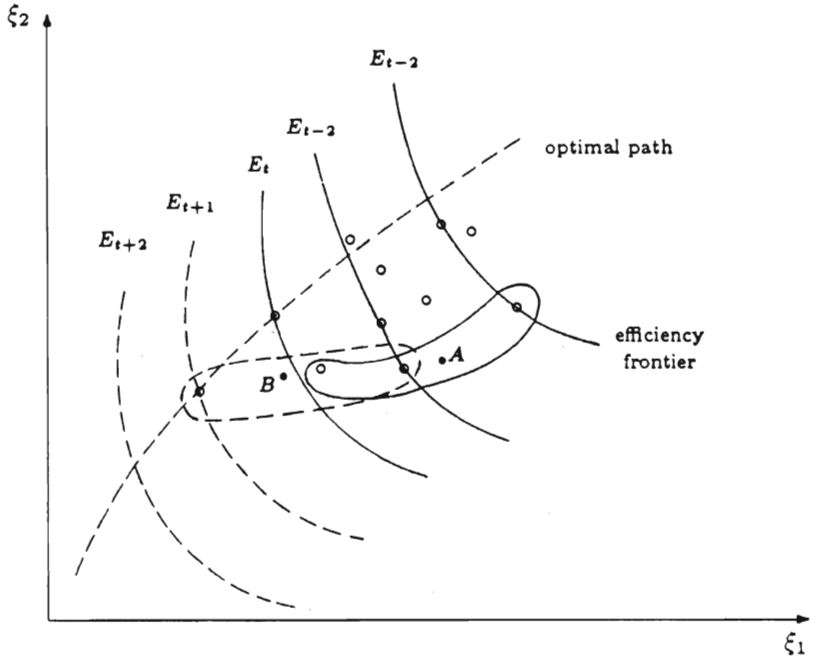


Figure 3.6: The process of structural change.

individual production units. One such firm, consisting of the units inside the solid line ellipse, is denoted by the closed circle A.

Initially, firm A comprised the production units inside the solid line ellipse. Let us assume that during the next period, $t + 1$, the oldest unit, situated at E_{t-2} is scrapped. Suppose at the same time a new production unit is built which is situated at the intersection between the optimal path and the efficiency frontier E_{t+1}. The firm will now consist of the units inside the broken line ellipse, and its centre of gravity will now move to B.

Vintage efficiency measures

In the vintage case it is difficult to find explicit measures of efficiency that are relevant from a policy point of view when the relevance of the measure is judged by the possibilities and the desirability of bringing the

structure closer to the frontier. Even if it is of limited help for policy purposes to look backward in time, estimating an optimal path for the industry and comparing this hypothetical optimal structural development with the actual one has a descriptive value. Particularly interesting is a comparison between the actual structure at a given moment and the hypothetical optimal one at the same moment, i.e., a comparison of two snapshot phases. This is the dynamic correspondence to the Farrell case. The measures we thus obtain will be called the Farrell vintage measures.

Let us assume that there exists an ex ante production function that is homogeneous of degree 1 with embodied technical progress, and that Figure 3.6 is applicable as an illustration of the development in this case. For an individual production unit, technical efficiency can be measured in the same way as in Figures 3.3 and 3.5. However, the relevant efficiency frontier for comparison is not the latest one, but the one existing at the respective investment date. This latter efficiency frontier shows the actual existing choice set for an investing firm at the time of investment. In the case of disembodied technical progress or learning-by-doing effects, the originally existing efficiency frontier should be adjusted. Such effects will shift the original frontiers towards the origin. With the same historical efficiency frontiers vintage price efficiency and vintage overall efficiency are obtained in the same way as in Figures 3.3 and 3.5.

If disembodied technical progress and learning by doing are allowed for when measuring price efficiency, no unique measure exists independent of the type of technical progress. We shall therefore disregard these effects. This strengthens the impression that price efficiency is a somewhat dubious concept. On the other hand, in the vintage case it becomes especially interesting to note whether or not the firms are satisfied with their original choice of factor proportions for the individual production units. The measure of technical efficiency now is interpreted as showing the firms' success in choosing capital equipment close to the relevant efficiency frontier. The measure of price efficiency shows the firm's success in forecasting the future factor price development and adaptation to it. Note that price efficiency does not mean adaptation to existing factor prices for the current inputs, but to the whole set of future factor prices during the life of the investment. The slope of the unit-cost line in Figure 3.3 is thus determined by an average of expected future factor prices.[14] The optimal path is derived from such an optimisation and reveals the optimal choice of factor proportions at any given moment. However, with discrete periods of time the optimal

[14] See Equations (2.4) and (2.5).

path is only defined for the intersections with the efficiency frontiers.

At the industry level, vintage technical and price efficiencies are obtained by weighing together the individual measures using the respective capacities. Since the measures for the individual units are relative, no problems arise when making a comparison between units belonging to different vintages. Overall vintage efficiency now shows the relative reduction in the amount of inputs needed to produce the observed output if the firms in the past had chosen production techniques from the efficient ex ante function existing at the respective investment dates, and if they at the same time had chosen factor proportions corresponding to the optimal path.

Even here it should be pointed out that the cost-minimising proportions are independent of the scale of production only in the case of homothetic production functions. Even in the case of non-homogeneous production functions the discussion on static efficiency is valid when efficiency is measured relative to the efficiency frontier corresponding to a given capacity at the investment date. However, when the development of demand is also taken into account, the notion of scale efficiency becomes different from that of the static case. Let us consider two possibilities:

1. *The technically optimal scale of the ex ante production function is relatively small compared to the increase in demand.*

In this case it is possible that the technically optimal scale may be realised. Normally, when the development of demand is continuous, investments will always be profitable if made at regular intervals over a period of time and only one new production unit is added at any given point of time (disregarding replacement investments). Even in the dynamic case a non-integer problem may arise, but probably only under rather special circumstances, such as an irregular development of demand or some kind of inertia in the planning process.

2. *The technically optimal scale of the ex ante production function is relatively large compared to the development of demand.*

In this case the *economically* optimal scale differs from the technically optimal scale and the capacities chosen for investments in new production units might correspond to the pre-optimal range of the production function, i.e., the elasticity of scale may be greater than 1. Thus, the efficiency measures based on *technically* optimal scale in Section 3.4 no longer apply directly.

To provide a concrete example, let us make the same assumptions as in Section 2.4. The ex ante production function is homogeneous of a degree

which is greater than 1 over its entire domain, and there is no technological progress. Demand grows at an exponential rate. The objective of the sector is to minimise the costs over the entire horizon, given the condition that capacity meets the demand at each point in time. The solution of this problem gives a sequence of optimal plant capacities, in spite of the fact that a technically optimal scale does not exist. In the input coefficient space there exists a set of efficiency frontiers not of different dates, but of different scales corresponding to the obtained sequence of optimal capacities. Figure 3.6 can be utilised even in this case.

In the case of embodied technological progress in the ex ante function there exists a set of efficiency frontiers both of different dates and scales. A measure of scale efficiency is obtained from the distance along a factor ray through the origin between the actual existing production unit and the efficiency frontier corresponding to the *economically* optimal scale. The production unit can be situated on either side of this frontier, since this frontier no longer delimits the technical possibilities, i.e., the comparison between the observed and the optimal input coefficients shows whether the capacity of the production unit is excessive or too small. One indication of structural scale efficiency may be obtained by weighing together the distances of the respective optimal capacities. Another measure, one easier to interpret, is the relation between the total costs of producing the given output with the existing units as compared to producing the same output with the hypothetically optimal units.

3.6 The characterisation of technical change

This section is devoted to a discussion of the characteristics of technical change. The impact of technical change may be measured in several ways, but here we will start with the measures introduced by Salter [1960]. We show how Salter's measure of technical advance may be generalised in a manner inspired by Farrell's decomposition of overall efficiency into technical and price efficiency.

Salter suggested three measures to describe technical advance:

(i) the *rate of technical advance* measured by the relative change in total unit cost for constant input prices and output level

(ii) labour, or capital saving *bias* measured by the relative change in the optimal (cost minimising) factor proportion for constant input prices

(iii) the relative change in the *elasticity of substitution.*

Assuming constant returns to scale, Salter considered only two factors. Here we generalise the first two measures to n factors in the case of non-homogeneous production functions. Generally, the relative change in cost for discrete time is

$$T = \bar{c}_{t+1}(x_{t+1}, q_1, \ldots, q_n)/\bar{c}_t(x_t, q_1, \ldots, q_n) \tag{3.33}$$

where $\bar{c}(\cdot)$ is the average cost function and q_i, $i = 1, \ldots, n$ are the factor prices equal for both periods. Salter compared unit costs for the same output level, i.e., $x_t = x_{t+1}$. When working with nonhomogeneous production functions it is natural to concentrate on the change in the minimum unit cost, i.e., x_{t+1} and x_t are the output levels that correspond to $\varepsilon_{t+1} = \varepsilon_t = 1$. This corresponds to the unit cost along the efficiency frontier in the input-coefficient space.

Generalised Salter measures

It might be of interest to note the similarity between this measure of technical advance and Farrell's [1957] concept of overall efficiency. In the two factor case this may be illustrated in the following way: Let P in Figure 3.7 be the point of reference on the efficiency frontier for the base period. Q' is the point on the efficiency frontier for a later period when factor prices remain the same. Assuming cost minimisation, a measure analogous to the Salter measure is the relative change in unit cost from P to Q', i.e., the unit cost reduction possible when choosing techniques from two different ex ante functions for constant factor prices and the achievement of optimal scale. This change is equal to OR/OP in Figure 3.7, which is also the Farrell overall efficiency measure with reference to the efficiency frontier at $t+1$ for a production unit with observed input coefficients given by P.

The Farrell overall measure can be split multiplicatively into technical efficiency, OQ/OP, and price efficiency, OR/OQ. When the factor ratio of the base period t is feasible in the next period $t+1$, the Salter technical advance measure can be split correspondingly. In our context this decomposition shows the relative reduction in unit cost due to the movement along a factor ray, T_1, and the movement along the next period efficiency frontier generated by biased technical change, T_2. Thus

$$T = T_1 \cdot T_2 \tag{3.34}$$

T_1 may be called *proportional technical advance* and T_2 *factor bias advance.*

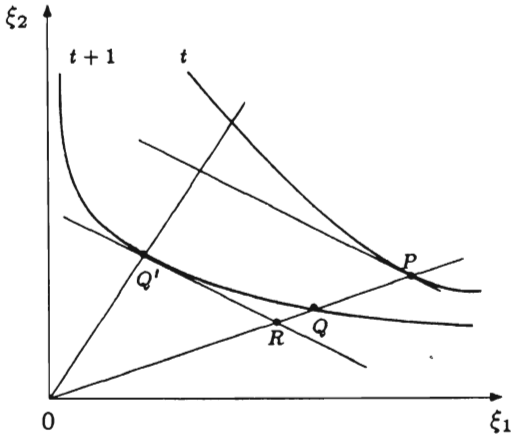


Figure 3.7: A decomposition of Salter's measure of technical advance.

Technical advance

As mentioned above, Salter compared unit costs for the same output level, i.e., $x_t = x_{t+1}$. He pointed out the lack of reference to economies of scale in the T measure, and suggested ways of measuring the impact of scale change on unit cost and factor bias. However, it might be preferable to make use of the relationship

$$\bar{c} = \varepsilon \partial c / \partial x = \varepsilon c'_x \tag{3.35}$$

where ε is the scale elasticity.[15] Insertion in (3.33) for $x_{t+1} = x_t = x$ yields

$$T = \left(\varepsilon_{t+1} \cdot c'_{x,t+1}(x, q_1, \ldots, q_n)\right) / \left(\varepsilon_t \cdot c'_{x,t}(x, q_1, \ldots, q_n)\right) \tag{3.36}$$

The change in unit cost is split up into the change due to the elasticity of scale changing and the change due to the change in marginal cost, for constant output and constant input prices.

When working with nonhomogeneous production functions it is natural to concentrate on the change in the minimum unit cost, i.e., when $\varepsilon = 1$. This corresponds to the unit cost along the efficiency frontier in the input-

[15] See Section 3.3.

coefficient space. From (3.36) we then have

$$T = c'_{x,t+1}(x^*_{t+1}, q_1, \ldots, q_n)/c'_{x,t}(x^*_t, q_1, \ldots, q_n) \tag{3.37}$$

where x^*_{t+1}, x^*_t are the output levels that correspond to $\varepsilon_{t+1} = \varepsilon_t = 1$.

Bias

The general version of the Salter bias measure is:

$$\begin{aligned} D_{ik} &= (v_{i,t+1}/v_{k,t+1})/(v_{i,t}/v_{k,t}) \\ &= \big[h_{i,t+1}(x_{t+1}, q_1, \ldots, q_n)/\big(h_{k,t+1}(x_{t+1}, q_1, \ldots, q_n)\big) \\ &\quad /\big(h_{i,t}(x_t, q_1, \ldots, q_n)/h_{k,t}(x_t, q_1, \ldots, q_n)\big)\big] \quad i,k = 1, \ldots, n \end{aligned} \tag{3.38}$$

where the $h(\cdot)$'s are the conditional factor demand functions and v_i, $i = 1$, $\ldots$, n, the inputs. It seems that Salter also assumed $x_{t+1} = x_t$.

In the case of more than two factors the Salter bias measure is a relative concept depending on the factor pair under consideration. If one wants a common basis for classifying the nature of bias, one possibility is to look at changes in the cost shares for constant input prices and output level. This has been proposed by Binswanger [1974], and also used by Stevenson [1980a], Greene [1983], and Kopp and Smith [1983].

To show the relationship between the Salter bias measures and the cost share measures we have the following expression for the change in the cost shares, C_i:

$$C_i = (q_i v_{i,t+1}/c_{t+1})/(q_i v_{i,t}/c_t) = \frac{v_{i,t+1}}{v_{i,t}} \cdot \frac{c_t}{c_{t+1}} \qquad i = 1, \ldots, n \tag{3.39}$$

Comparing this expression with Equation (3.38), the Salter bias measure may be interpreted as the relative change of factor No. i between the two points in time, weighted for the relative change of the other factor under consideration, measured in the opposite direction of time. In the cost share measure the relative change in average or total costs is substituted as an index weight, thus constituting a common weight applied to all factors.

3.7 Concluding remarks

Efficiency is a word that is easy to use, but difficult to give a precise operational meaning. The efficiency measures reviewed in this chapter

are best suited as descriptions of the structure of establishments within industries. The interpretation of efficiency measures essentially depends on the specification of production structure, such as scale properties and rigidity of factor proportions ex post.

If somewhat more realistic assumptions than those usually made are allowed, measures of price efficiency are soon unmanageable from an interpretative point of view.

As discussed in Chapter 2, a word of caution is warranted with regard to the normative use of efficiency measures. Efficiency measures provide a description of the structure of an industry and a necessary step for identifying the causes of efficiency differences. But if, for instance, capital clayishness is the cause of efficiency differences, it is not economically relevant to pursue a policy of bringing all units up to the standard of the most efficient vintage. Differences in measured efficiency might correspond to differences in the age of equipment, and this tells us nothing about the *economic* efficiency of the equipment. The point then is to optimise an ongoing process of structural change. Thus, it may not be relevant to use the frontier function "to ascertain the maximum productive capacity of an industry" (Aigner and Chu [1968], p. 830).

From a policy point of view comparisons between best-practice establishments and the industry average provide a valuable description of the structure. It must again be stressed that differences in efficiency are not necessarily undesirable. In a putty-clay world the policymaker must take the optimal path of structural development as a reference for action. The policy problem is to implement this path, directly or indirectly, influencing the rate and direction of new investments and scrapping of old units.

Appendix 3.1: Further aspects of the efficiency frontier

Since the efficiency frontier is a central concept in the derivation of efficiency measures, it warrants a more detailed exposition. Let us consider equation (3.4),

$$\varepsilon = \Phi(v_1, \ldots, v_n) \tag{A3.1}$$

From the definition of a regular ultra-passum law, we must have

$$\frac{\partial \varepsilon}{\partial v_i} = \Phi_i' < 0 \qquad i = 1, 2, \ldots, n \tag{A3.2}$$

In the two-factor case, the following expression applies to the slope of the contour lines of the passus coefficient

$$\frac{dv_2}{dv_1} = -\frac{\Phi'_1}{\Phi'_2} < 0 \qquad \text{(A3.3)}$$

This means that for regular ultra-passum laws the optimal scale curve is a falling curve in the factor diagram. This is the only restriction implied by our class of production laws.

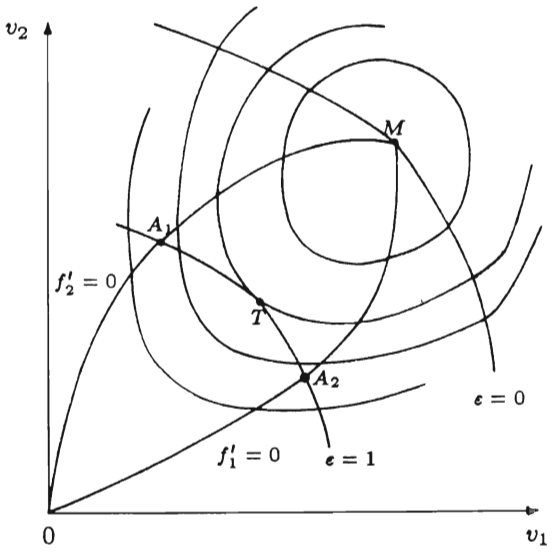
The contour curve obtained for $\varepsilon = 1$, the optimal scale curve, generally intersects some isoquants. From the transformation of the optimal scale curve to the input-coefficient space it follows directly that the intersection point must be on the efficiency frontier. Thus, in general, output is not constant along the efficiency frontier. The shape of the transformed isoquants is not obvious, but the isoquants intersected by the optimal scale curve must necessarily intersect in the input-coefficient space.

The situation may be illustrated in Figure A3.1 for the standard case of a regular product surface with a maximum point for finite values of the factor quantities and isoquants that are convex to the origin.

The region confined by OA_1MA_2O is the substitution region defined as the region where the marginal productivities are non-negative. The point M represents the global maximum point of production. (Of course, the contour line for $\varepsilon = 0$, the technically maximal scale, runs through this point.) We can distinguish between three cases with regard to the form of the substitution region and the situation of the optimal scale curve:

(i) the optimal scale curve passes through the substitution region and goes out in the factor space on both sides
(ii) the optimal scale curve is outside the substitution region only on one side
(iii) the complete range of the optimal scale curve lies inside the substitution region.

In the general case (i) the tangents to the isoquants in the two-factor case are horizontal on the lower boundary ($f'_1 = 0$) of the substitution region, and vertical on the upper boundary ($f'_2 = 0$). According to the property of the regular ultra-passum law the optimal scale curve must be falling over its entire range. This means that the curve must have a tangency point with an isoquant inside the substitution region. This tangency point (T in the figure) represents the maximum quantity that can be produced in the technically optimal scale. With isoquants that are convex to the origin and an optimal scale curve which is either concave or convex, it can be seen



The region of substitution and the optimal ($\varepsilon = 1$) and the maximal ($\varepsilon = 0$) scale curves

Figure A3.1: An isoquant map.

that this tangency point, corresponding to maximum output, is unique. (These curvatures are, however, not implied by the sufficient second-order conditions.)

The diagram also shows that all the isoquants representing production quantities between zero and the maximum on the optimal curve are (according to the assumption about the curvature of the optimal scale and with $f(v_1, v_2, \ldots, v_n) = 0$ for $v_i = 0$, $i = 1, 2, \ldots, n$) intersected twice by the curve of optimal scale.

The first-order condition (3.7) for maximum values of the average productivities under proportional variation may also be written (inserting

$$\frac{df}{d\mu} = \sum_{k=1}^{n} \frac{\partial f}{\partial v_k} \frac{dv_k}{d\mu}$$

in the second expression in (3.7) and rearranging)

$$\frac{f}{v_i} = \frac{\partial f}{\partial v_i} + \sum_{k \neq i} \frac{\partial f}{\partial v_k} \frac{dv_k}{d\mu} \Big/ \frac{dv_i}{d\mu} \qquad i,k = 1,2,\ldots,n \tag{A3.4}$$

Maximising the average productivity of factor No. i without any restriction yields the necessary first-order conditions

$$\begin{aligned} &\frac{\partial(f/v_i)}{\partial v_i} = \frac{v_i \partial f/\partial v_i - f}{v_i^2} = 0 \Rightarrow \frac{f}{v_i} = \frac{\partial f}{\partial v_i} && i = 1,2,\ldots,n \\ &\frac{\partial(f/v_i)}{\partial v_k} = \frac{\partial f/\partial v_k}{v_i} = 0 \Rightarrow \frac{\partial f}{\partial v_k} = 0 && k \neq i \end{aligned} \tag{A3.5}$$

By comparing (A3.4) and A3.5) we see that when $\partial f/\partial v_k = 0$ ($k = 1$, $\ldots$, $i-1$, $i+1$, $\ldots$, n) the maximum value of the average productivity of input i under proportional variation is identical with the unconstrained maximum value. In Figure A3.1 this point is represented by the intersection point A_2 ($i = 1$) or A_1 ($i = 2$) between the curve of optimal scale and the boundary of the substitution region.

When the curve of optimal scale lies inside the substitution region (case (iii)), free maximum values of the average productivities are absent.

Transforming the isoquant map in Figure A3.1 yields transformed isoquants with the shape seen in Figure A3.2.

The points A_1' and A_2' in Figure A3.2 correspond to the points A_1 and A_2 in Figure A3.1 and represent the global minimum values of the unit requirements. The part of the technique line between A_1' and A_2' is the part inside the substitution region. The whole border of the technically feasible region will be inside the substitution region in case (iii), i.e., the curve of optimal scale lies inside the economic region.

The efficiency frontier in Figure A3.2 is drawn convex to the origin. The only restriction on the curvature is that a ray passing through the origin can only have one point in common with the efficiency frontier. Applying the assumption that isoquants are convex to the origin, the part of the efficiency frontier which lies between A_1' and A_2' must be a falling curve. This is so, since, in the general case with the efficiency frontier as an envelope to the transformed isoquants, every point on the line is a tangency point with an isoquant (i.e., the line has the same slope as an isoquant). The tangents to the efficiency frontier will pass asymptotically through the origin when going outwards from A_1' and A_2'. In case (iii) the efficiency frontier will run asymptotically to the lines parallel to the axes, representing the asymptotic minimum values of the unit requirements.

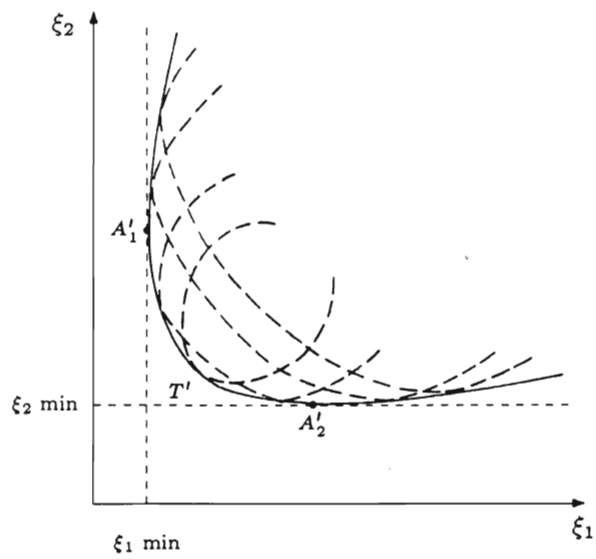


Figure A3.2: The feasible region in the input-coefficient space.

It might be assumed that convex isoquants are all that is required to ensure a convex efficiency frontier, but even with the additional restriction of a regular ultra-passum law it can be shown that this is not the case.

A sufficient additional restriction is that the optimal scale curve has only one point of intersection with each isocline, i.e., the geometric locus for points with a constant rate of marginal substitution.

4

Empirical Approaches to the Frontier Production Function

4.1 Introduction

The recent interest in frontier production functions has as its starting point the seminal work of Farrell [1957] on how to measure productive efficiency. As discussed in Section 3.4 his frame of reference for efficiency measures is the convex hull of the observed input-coefficients (unit requirements) in the input coefficient space, assuming that the unspecified frontier function is homogeneous of degree 1. Such a convex hull is called an efficient isoquant. When it is assumed that the industry frontier function exhibits increasing returns to scale, there does not exist a unique frame of reference for efficiency measures.[1] In Farrell and Fieldhouse [1962] an efficient isoquant is constructed for each chosen level of output so as to serve as frames of reference for efficiency measures.

The convex hull may be regarded as a pessimistic estimate of the *efficiency frontier*[2] of the underlying frontier function. The efficiency frontier constitutes the boundary towards the axes of the technically feasible region in the input-coefficient space, and is the locus of points where the elasticity of scale equals 1.[3]

One advantage of Farrell's method is that it is easy to apply when the underlying function is linearly homogeneous.[4] The considerably more cumbersome method of computing an efficient isoquant for chosen output levels for the increasing returns to scale is used in Seitz [1970, 1971].

---

[1] See Section 3.4.

[2] This concept was introduced in Section 3.3.

[3] See Section 3.3 and Appendix 3.1.

[4] See, e.g., Todd [1971, 1985] and Meller [1976].

In the case of non-homogeneous production functions it is obviously more advantageous to use an explicitly specified function. An estimate of the efficiency frontier, even when it is smoothly curved, yields insufficient information for the establishment of a production function except in the case of constant returns to scale. It is necessary to have an explicit function in order to compute the complete set of efficiency measures. [5]

4.2 Estimation of parametric frontier production functions

In Section 3.2 the frontier function to be estimated was defined as yielding maximum output for a given level of inputs. It must then bound observed outputs from above. Each unit of observation can be represented by

$$x^j = f(v^j)e^j, \quad e^j \in (0,1] \qquad j = 1,\ldots,N \tag{4.1}$$

where $f(v)$ is the industry frontier function, v a vector of inputs, e^j the output increasing efficiency measure associated with each unit No. j, and N the number of units.

Since firms' performances may be affected by factors entirely outside their control (such as poor machine performance, bad weather, input supply disruptions, various kinds of breakdowns, etc.), it may also be realistic to allow observations to be *above* the frontier.

The frontier is called *deterministic* if all the observations must lie on, or below the frontier, and *stochastic* if observations can be above the frontier due to random events.[6]

These two basic approaches put natural restrictions on the estimation procedures. In addition, the assumptions about the distribution of the efficiency variable e pose additional restrictions.

The different approaches are summarised in Table 4.1. The elimination of inefficient observations and the estimation of an average function on the remaining sample, as in Kurz and Manne [1963], falls outside this scheme. The same holds for the successive removals of the frontier observations, as done by Timmer [1971].

---

[5] See, however, Grosskopf [1986] for a discussion of the advantages of the flexibility of the non-parametric programming approach.

[6] See Aigner et al. [1977] and the surveys in Førsund et al. [1980] and Schmidt [1985–86].

Table 4.1: A survey of original approaches for the estimation of explicit frontier production functions.

Stochastic specification	Types of frontiers	
	Deterministic frontier: Entire sample on or below the frontier	Stochastic frontier: No "on or below the frontier" restrictions on observations
No explicit efficiency distribution	Programming methods: Aigner and Chu [1968], Førsund and Hjalmarsson [1979a]	
Explicit efficiency distribution	Maximum likelihood programming methods: Schmidt [1976], Broeck et al. [1980]	Corrected ordinary least squares: Richmond [1974]
Specific random distribution with different weights on positive and negative residuals		Maximum likelihood: Aigner et al. [1976]
Composed error: explicit efficiency and random distributions		Maximum likelihood: Aigner et al. [1977], Meeusen and Broeck [1977a]

In order to illustrate the differences between the deterministic and the stochastic approach, the typical relative positions of the graphs of the estimated production functions are illustrated in Figure 4.1. One standard criticism of frontier functions determined from observed data is that "outliers" have too much influence on the resulting frontier. However, for the "true" frontier, efficient outliers *should* in principle count disproportionately. The approach in Timmer [1971] of estimating a so-called probabilistic frontier by removing the efficient observations on the frontier and then

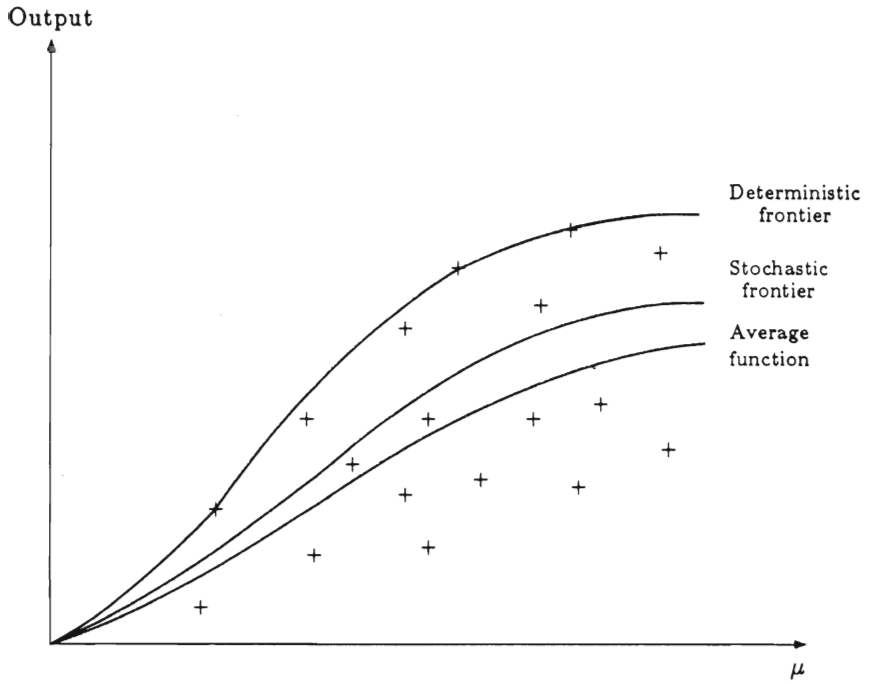


Figure 4.1: An illustration of the typical position of deterministic and stochastic frontier productions function and the traditional average functions.

recomputing a deterministic frontier appears too arbitrary. The stochastic frontier seems a more appropriate answer to the outlier "problem". It is a real problem if the arguments for introducing the purely random term are relevant. Aigner et al. [1977, p. 25] point to "external events such as luck, climate, topography, and machine performance", and errors of observation and measurement. But all the variables of the first set seem, in principle, observable[7] and may therefore be entered as explanations of differences in economic performance. As stated in the introduction, one of the objectives of the frontier function is to serve as a basis for the identification of explanatory factors. Measurement errors appear to be unobservable, so here it is a question of what information we have about the quality of the data.

[7] Luck must have a quantitative measure.

4.3 Deterministic frontier

Deterministic frontiers without an explicit efficiency distribution

Aigner and Chu [1968] provided a framework for computing an explicit production function of the Cobb-Douglas type taking into account (unrestricted scale elasticity) the restriction that the observations should be on, or below the function. Their point of departure was that this frontier function represents the correct conceptual construct from the core of microeconomic theory i.e., it yields maximum output from given inputs.[8] This may also be said to be the case for Afriat [1972]. However, in a cross section sample of production units, each unit may be perfectly efficient within its own technology. If, for instance, putty-clay is a valid assumption about production structure,[9] the notion of a frontier function shared by all firms is unnecessarily restrictive. It should be noted that the estimating of a single frontier $f(v)$ as in (4.1) by utilising a sample of individual units is not the same as assuming that the observations are generated by this frontier. The frontier function is the most efficient function the data can support.

Afriat's approach is based on representing the operations of the units as efficient, or if not exactly so, then as close to efficient as possible. Thus his frontier function is based on maximising an increasing function of output-efficiency measures (i.e., observed outputs compared with potential outputs on the frontier), which means that the objective is to get the observations "as close as possible" to the frontier in the output direction. The objective functions of simple and squared deviations from the frontier in the output dimension in Aigner and Chu [1968] are examples of this approach. In general, the distance from the individual observation to the frontier can be measured in several ways. Measuring the distance in the output direction implies that output increasing efficiency measure E_2 in Chapter 3 is the main concern. An alternative would be to measure the distance in the direction corresponding to input saving efficiency measure E_1.

We will generalise the approach in Aigner and Chu [1968] to allow for variable returns to scale. The frontier function is prespecified to be a homothetic function of the general form:

$$G(x) = g(v) \tag{4.2}$$

[8] See also Aigner et al. [1977].

[9] See Chapter 2.

where x = rate of output, v = vector of inputs, $G(x) = a$ a monotonically increasing function and $g(v) = a$ a homogeneous function of degree 1.

The function is fitted in such a way that the following relationship holds for each unit

$$x = G^{-1}(g(v))e, \qquad e \in (0,1] \tag{4.3}$$

When parametricising $G(\cdot)$ it turns out to be more convenient to measure the distance from the frontier in the direction corresponding to input saving:

$$G(x) = g(v) \cdot u \tag{4.4}$$

where obviously $u \in (0,1]$.

The transformation function $G(\cdot)$ is specified in the following way[10]:

$$\ln G(x) = \alpha \ln x + \beta x \tag{4.5}$$

The function $g(\cdot)$ is specified as a Cobb-Douglas (C-D) function. For computational convenience the following increasing function in the efficiency measures $\ln u^j$ is to be maximised:

$$\sum_{j=1}^{N} \ln u^j = \sum_{j=1}^{N} \left(\alpha \ln x^j + \beta x^j - \ln A - \sum_i a_i \ln v_i^j \right) \tag{4.6}$$

subject to the "on-or-below-the-frontier" constraints

$$\alpha \ln x^j + \beta x^j - \ln A - \sum_i a_i \ln v_i^j, \leq 0 \qquad j = 1, \ldots, N \tag{4.7}$$

and the homogeneity constraint

$$\sum_i a_i = 1 \tag{4.8}$$

The computational procedure thus implies the solution of a standard LP-problem.

With regard to functional forms the available workable production functions may also be employed as frontier functions. So far, the Cobb-Douglas function has been the most popular.[11] A homothetic function with

[10] Cf. Zellner and Revankar [1969].

[11] See Aigner and Chu [1968], Timmer [1971] and Carlsson [1972].

a C-D kernel function permitting variable elasticity of scale has been estimated in Førsund and Hjalmarsson [1979a]. As mentioned by Afriat [1972, p. 568], the special restrictive properties of these functions "are not deliberate empirical hypotheses, but are accidental to technical convenience of the functions".

It is possible to maintain the LP computational framework even for more general functional forms such as translog:

$$\ln x = \ln a_0 + \sum_i a_i \ln v_i + \tfrac{1}{2} \sum_i \sum_k \gamma_{ik} \ln v_i \ln v_k + \ln e \qquad (4.9)$$

By substituting $\ln e$ for $\ln u$ in (4.6) and accordingly changing the constraints in (4.7) we are still left with a LP problem. Various parameter restrictions can be substituted for (4.8). Since translog is an explicit form in output, we have now returned to the original output deviation measure, e, introduced in (4.1).

Deterministic frontiers with an explicit efficiency distribution

This approach is based on the assumption that the frontier function can be inferred from a probabilistic hypothesis about efficiencies. Given the functional form of the frontier function and the probability distribution of the efficiency variable, the parameters of both functions and estimated measures $e^j = x^j / f(v^j)$, $j = 1, \ldots, N$, have maximum likelihood.[12]

Afriat suggested the beta distribution as the most general probability distribution satisfying the requirements of such an efficiency distribution. Ideally the efficiency distribution should be derived from the economic mechanism generating the efficiency differences between units. But it might be too difficult to identify and model such basically dynamic mechanisms within an explicit efficiency distribution for a cross section of units.[13]

If it is possible to establish an explicit efficiency distribution and a specific functional form of the frontier function, then it is natural to derive maximum likelihood estimates of the parameters of the frontier function and the efficiency distribution.[14]

A basic problem with such ML estimators, as pointed out in Schmidt [1976], is that due to the "on-or-below-frontier" constraints a regularity

[12] See Afriat [1972], p. 581.

[13] See Chapter 2 for analyses of such mechanisms.

[14] For a discussion and application of this approach see Gabrielsen [1975], Schmidt [1976, 1978], Chu [1978], and Førsund and Jansen [1977].

condition for the application of maximum likelihood is not met. This is due to the fact that the range of the stochastic output variable depends on the parameters to be estimated. It is not known whether these ML estimators are consistent and asymptotically efficient. In Greene [1980a] it is shown that the desirable asymptotic properties still hold if the density of $\ln e$ satisfies the conditions that it is zero at $\ln e = 0$ and the derivative of the density of $\ln e$ with respect to its parameters approaches zero as $\ln e$ approaches zero.

As noted by Greene, the gamma density satisfies this criterion and is thus potentially useful here. However, it is a little troubling that one's assumption about the distribution of technical inefficiency should be governed by statistical convenience.

As an example we shall again assume that the frontier production function is of the general homothetic form of (4.2). The efficiency variable u is now interpreted as stochastic, implying input-neutral differences between units with respect to what they get out of their inputs. The inputs are assumed to be exogenous and u is assumed to be identically and independently distributed. It is convenient to consider (4.4) in logarithmic form.

The joint log-likelihood function for the output variables on the left-hand side of (4.4) is:

$$\begin{aligned} \ln \ell(x^1, \ldots, x^N) &= \sum_{j=1}^{N} \ln h(\ln u^j) + \ln |J| \\ &= \sum_{j=1}^{n} \ln h\big(\ln G(x^j) - \ln g(v^j)\big) \\ &\quad + \sum_{j=1}^{N} \ln |\partial \ln G(x^j)/\partial x^j| \end{aligned} \tag{4.10}$$

where $h(\cdot)$ is the distribution function for $\ln u$, and the second term J is the Jacobian determinant due to the implicit form of the production function.

We now insert specific functional forms which enable us to derive ML-estimators. The following one-parameter distribution will be used

$$h(\ln u) = (1+a)e^{(1+a)\ln u}, \quad a > -1, \quad \ln u \in (-\infty, 0] \tag{4.11}$$

$$E(\ln u) = \frac{1}{1+a}, \qquad \text{Var}\,(\ln u) = \frac{1}{(1+a)^2} \tag{4.12}$$

From (4.11) the distribution of $u, k(u)$ follows directly,

$$k(u) = h(\ln u)\left|\frac{\partial \ln u}{\partial u}\right| = (1+a)\cdot u^{1+a}\left|\frac{1}{u}\right| = (1+a)u^a \tag{4.13}$$

Since u is identical to our input saving measure E_1, we are interested in the expected value of u:

$$E(u) = \frac{1+a}{2+a} \tag{4.14}$$

Inserting (4.11) in (4.10) yields

$$\ln \ell(x^1,\ldots,x^N) = N\ln(1+a) + (1+a)\cdot\sum_{j=1}^{N}\left(\ln G(x^j) - \ln g(v^j)\right)a + \sum_{j=1}^{N}\ln\left|\partial \ln G(x^j)/\partial x^j\right| \tag{4.15}$$

Maximising $\ln \ell(\cdot)$ with respect to a and putting the derivative equal to zero yields

$$\frac{N}{1+a} + \sum_{j=1}^{N}\left(\ln G(x^j) - \ln g(v^j)\right) = 0 \tag{4.16}$$

If ML-estimates for the production function parameters were available, an ML-estimate for a, $\hat{a}$, is obtained by rearranging (4.16),

$$\hat{a} = \frac{N}{-\sum_{j=1}^{N}\left(\ln G(x^j) - \ln g(v^j)\right)} - 1 \tag{4.17}$$

(4.17) means that an ML-estimator for a is derived by using the average value of $\ln u^j$

$$\left[\frac{1}{N}\sum_{j=1}^{N}\left(\ln G(x^j) - \ln g(v^j)\right)\right]$$

as an estimate for the expected value, keeping in mind (4.12).

Eliminating a in (4.15) by inserting (4.17) in (4.15) yields the concen-

trated log-likelihood function:

$$\ln \ell^* = N \ln N - N + N \ln \left[\sum_{j=1}^{N} \left(\ln G(x^j) - \ln g(v^j) \right) \right] + \sum_{j=1}^{N} \ln \left| \partial \ln G(x^j)/\partial x^j \right| \tag{4.18}$$

By using (4.18) as the objective function it is possible to proceed to derive the estimates of the parameters of the $G(\cdot)$ and $g(\cdot)$ functions by maximising (4.18), subject to the on-or-below frontier constraints,

$$\ln G(x^j) - \ln g(v^j) \leq 0 \qquad j = 1, \ldots, N \tag{4.19}$$

and the homogeneity constraint on $g(\cdot)$. The sum of the slacks in (4.19) may now be inserted in (4.17) yielding a ML-estimate of a.

Employing the same functional forms used in (4.6) yields a non-linear objective function and the same linear constraint set as in (4.7). The objective function has the following, non-linear form:

$$\ln \ell^* = N \ln N - N - N \ln \left(\sum_{j=1}^{N} (\alpha \ln x^j + \beta x^j) - \ln A - \sum_i a_i \ln v_i^j \right) + \sum_{j=1}^{N} \ln \left| \beta + \frac{\alpha}{x^j} \right| \tag{4.20}$$

4.4 Stochastic frontiers

Utilising average functions when estimating frontier functions

One approach to estimating frontier functions is to utilise the "average" function parameters estimated by standard regression techniques (such as ordinary least squares, OLS) except for the constant term or level of the function, the estimation of which is adapted especially to conform to frontier function restrictions. This approach was apparently first noted by Richmond [1974]. We will call this *corrected* OLS, or COLS. Suppose for

simplicity that (4.1) is linear as in the Cobb-Douglas function. Then if we let μ be the mean of $\ln e$, we can write

$$\ln x = (\ln A + \mu) + \sum_{i=1}^{n} a_i \ln v_i + (\ln e - \mu) \tag{4.21}$$

where the new error term has zero mean. Indeed the error term satisfies all of the usual ideal conditions except normality. Therefore, (4.21) may now be estimated by OLS to obtain best linear unbiased estimates of $(\ln A + \mu)$ and of the a_i. If a specific distribution is assumed for $\ln e$ and if the parameters of this distribution can be derived from its higher-order (second, third, etc.) central moments, then we can estimate these parameters consistently from the moments of the OLS residuals. Since μ is a function of these parameters, it too can be estimated consistently, and this estimate can be used to "correct" the OLS constant term, which is a consistent estimate of $(\ln A + \mu)$. COLS thus provides consistent estimates of all of the parameters of the frontier.

A difficulty with the COLS technique is that, even after correcting the constant term, some of the residuals may still have the "wrong" sign so that these observations end up above the estimated production frontier. This makes the COLS frontier a somewhat awkward basis for computing the technical efficiency of individual observations. One response to this problem is provided by the stochastic frontier approach discussed below. Another way of resolving the problem is to estimate (4.21) by OLS, and then to correct the constant term not as above, but rather by shifting it up until no residual is positive and one residual is zero. Gabrielsen [1975] and Greene [1980a] have both shown that this correction provides a consistent estimate of $\ln A$.

Another difficulty with the COLS technique is that the correction of the constant term is not independent of the distribution assumed for $\ln e$. Consider the one-parameter gamma distribution

$$h(\ln e; \sigma) = \frac{1}{\Gamma(\sigma)}(-\ln e)^{\sigma-1} \exp(\ln e), \quad -\infty < \ln e < 0, \quad \sigma > 0 \tag{4.22}$$

The first two moments are $E(\ln e) = \text{var}\ (\ln e) = -\sigma$. Hence the OLS-variance estimator provides the correction to the constant term.

$$\text{Var}\ (\ln e) = \frac{1}{N-k-1} \sum_{j=1}^{N} \left[\ln x^j - \left(\ln A + E(\ln e) \right) - \sum_{i=1}^{n} a_i \ln v_i^j \right]^2 \tag{4.23}$$

Now consider the exponential distribution

$$h(\ln e;\sigma) = \frac{1}{\sigma}\exp(\ln e/\sigma), \qquad -\infty < \ln e < 0, \qquad \sigma \neq 0 \tag{4.24}$$

where $\sigma = 1/(1+a)$ as in (4.12), with the first two moments $E(\ln e) = -\sigma$ and Var $(\ln e) = \sigma^2$. Hence the negative *square root* of the OLS variance estimator provides the correction to the constant term. Thus the one-parameter gamma distribution and the exponential distribution yield systematically different corrections for the constant term, and systematically different estimates of technical efficiency, except for the special case Var $(\ln e) = 1$. For example, Richmond's applications of the results in Griliches and Ringstad [1971] revealed quite high estimates of technical efficiency for Norwegian manufacturing. Recomputing these using the exponential distribution, mean efficiency falls from 87% to 69%. Note that this problem does not arise if the constant term is estimated by shifting the function upward, as just described.

One general disadvantage of using the form of the average function as a kind of "lid" on, or above, the observations is that the difference between the average and the frontier function is only allowed to be expressed by the constant term or level parameter. This precludes the discovery of possibly interesting differences regarding, e.g., marginal productivities of the inputs.

Stochastic frontiers with a composed error specification

In the preceding sections all variation in firm performance is expressed by variations in firm efficiencies relative to the common frontier. Sometimes this proves difficult to justify. It may be empirically relevant that a firm's performance may be affected by factors entirely outside its control (such as poor machine performance, bad weather, input supply breakdowns, and so on), as well as by factors under its control (such as inefficiency). To lump the effects of exogenous shocks, both fortunate and unfortunate, together with the effects of measurement error and inefficiency into a single one-sided error term, and then to label the mixture "inefficiency" may be somewhat questionable.

This conclusion is reinforced if one also considers the statistical "noise" that every empirical relationship contains. The standard interpretation is that first, there may be measurement error (hopefully on the dependent variable and not on the independent variables). Second, the equation may not be completely specified (hopefully with the omitted variables individually unimportant). Both of these arguments hold just as well for production

functions as for any other kind of equation, and it is dubious at best not to distinguish this "noise" from inefficiency, or to assume that "noise" is one-sided. Aigner et al. [1976] countered this problem by allowing observations to be above the frontier, but placed different weights on positive and negative disturbances. This approach was more satisfactorily developed in Aigner et al. [1977] and Meeusen and van den Broeck [1977a].

The essential idea behind the stochastic frontier model is that the error term is composed of two parts. A symmetric component permits random variation of the frontier across firms and captures the effects of measurement error, other statistical "noise" and random shocks outside the firm's control. A one-sided component captures the effects of inefficiency relative to the stochastic frontier. A stochastic production frontier model may be written as

$$x = f(v)e^{-\theta} \cdot e^{-\omega} \tag{4.25}$$

where the stochastic production frontier is $f(v) \cdot e^{-\theta}$, and $e^{-\theta}$ has a symmetric distribution, i.e., $e^{-\theta} \in (0, \infty)$ and $E(e^{-\theta}) = 1$, so as to capture the random effects of measurement error and exogenous shocks that cause the placement of the deterministic kernel $f(v)$ to vary across firms. Technical inefficiency relative to the stochastic production frontier is then captured by the one-sided error component $e^{-\omega}$, $e^{-\omega} \in (0, 1]$.

The range of $e^{-\omega}$ ensures that all observations lie on or beneath the stochastic production frontier. Unfortunately there is no fully satisfactory way of determining whether the observed performance of a particular observation, compared with the deterministic kernel of the frontier, is due to inefficiency or to random variation in the frontier. This constitutes the main weakness of the stochastic frontier model: it is not possible to exactly decompose individual residuals into their two components, and so it is not possible to get technical inefficiency measures for an individual observation. However, one can obtain an estimate of the mean efficiency over the sample.

One way of obtaining information on individual efficiencies is proposed by Jondrow et al. [1982]. Taking logarithms in (4.25) and considering the expected value of ω^j, conditional on $(\omega^j + \theta^j)$, information is obtained about ω^j for each unit. This conditional distribution contains whatever information $\omega^j + \theta^j$ yields about ω^j. Either the mean or the mode of this distribution can be used as a point estimate of ω^j. The remaining shortcoming of this decomposition is that these estimates of individual efficiencies are not consistent, i.e., the variability intrinsic to the conditional distribution $(\omega^j; \omega^j + \theta^j)$ is independent of sample size. As pointed out in

Jondrow et al. [1982] this reflects the obvious fact that $(\omega^j + \theta^j)$ contains only imperfect information about ω^j.

Direct estimates of the stochastic production frontier model may be obtained by either maximum likelihood or COLS methods. Introducing probability distributions for θ and ω, assuming that θ and ω are independent and that x is exogenous, the asymptotic properties of the maximum likelihood estimators can be proved in the usual way. [15] When expressed in linear form, COLS may also be used to estimate the model by adjusting the constant term by the appropriate function $E(\omega)$, which is derived from the moments of the OLS residuals. The COLS estimates are easier to compute than the maximum likelihood estimates, although they are asymptotically less efficient. Olson et al. [1980] present Monte Carlo evidence indicating that COLS generally performs as well as maximum likelihood, even for rather large sample sizes.

Whether the model is estimated by maximum likelihood or by COLS, the distribution of ω must be specified. Aigner et al. [1977] and Meeusen and van den Broeck [1977] considered half-normal and exponential distributions, respectively, for ω. Both of these distributions have a mode of zero. Stevenson [1980b] has shown how the half-normal and exponential distributions can be generalised to truncated normal and gamma, respectively. Both of these generalisations can have non-zero modes, with zero modes being testable special cases.

Stochastic frontier models have been applied to a variety of data sets, including data om Brazilian manufacturing, Columbian enterprises, the Indonesian weaving industry, U.S. steam electric generating plants, the U.S. primary metals industry, U.S. agriculture, French and Yugoslavian manufacturing and Finnish breweries.[16] Data on milk processing in Swedish dairy plants have also been examined,[17] and in Chapter 7 MLE estimates of a stochastic production frontier model are compared with two sets of estimates of a deterministic production frontier model.

---

[15] Note that the presence of the symmetric error component θ solves the bounded-range problem encountered by some variants of the deterministic frontier model.

[16] For the study on Brazilian manufacturing, see Lee and Tyler [1978]; for the study on Columbian enterprise data, see Tyler and Lee [1979]; and for the study of the Indonesian weaving industry, see Pitt and Lee [1981]. The U.S. steam electric generating plants were examined in Kopp and Smith [1978], the U.S. primary metals industry in Aigner et al. [1977] and U.S. Agriculture in Aigner et al. [1976]. French manufacturing was studied in Meeusen and van den Broeck [1977b]. Yugoslavian manufacturing sectors were studied in Nishimizu and Page [1982]. Finnish breweries were studied in Summa [1986].

[17] See Broeck et al. [1980].

Panel estimation

The database in the Indonesian weaving industry constitutes a panel, and utilising a variance components model approach it was possible to test whether the efficiency variable is constant over time for each unit. Maximum likelihood estimates of a model with a Cobb-Douglas specification and a time invariant efficiency component yielded mean efficiency of 60 to 70 per cent, and in contrast to other studies, the variance of the efficiency variable was not swamped by the variance of the random variable. Testing the time invariance assumption indicated that the most appropriate model would be one permitting efficiency to vary over time for some units.

The variance components model has been further analysed in Schmidt and Sickles [1984]. By utilising the ideal behind corrected least squares, that is, correcting the constant term (within a linear production model, e.g., Cobb-Douglas on logarithmic form), they demonstrate how *individual* measures of efficiency can be obtained if it is assumed that the efficiency component is time invariant. As previously pointed out the location of the frontier is determined by using the largest estimated individual constant term as an estimate of the frontier function constant. The time invariance assumption makes decomposition into the efficiency term and the random term possible. In contrast to the conditional estimators of efficiency discussed above the estimators within the variance components model are consistent.

Estimation of CE-models

We shall now consider in more detail the estimation of stochastic frontier functions. The maximum likelihood method may be applied if it is assumed that the inputs and θ and ω are mutually independent. Suppose that the estimation is on logarithmic form, a form which is the most convenient analytically. The first step is to find the joint distribution for $\ln\omega + \ln\theta$. In order to obtain an estimate of the average level of efficiency, e.g., $E(\omega)$, the distribution for ω must be such that there is a unique relation between the expected value and the variance.[18]

Considering the general homothetic function used in Section 4.3, the concept of a stochastic frontier emerges when the variable u in (4.4) is multiplicatively decomposed into one pure random term e^{-z_0} and one sys-

[18] We have that $E(\ln\omega+\ln\theta) = E(\ln\omega)$ and Var $(\ln\omega+\ln\theta)$ = Var $(\ln\omega)$+Var $(\ln\theta)$.

tematic term e^{-z_1} distributed in the interval $(0, 1]$,

$$u = e^{-z_0 - z_1} \tag{4.26}$$

It is natural to assume that z_0 is normally distributed, $N(0, \sigma)$. If z_1 is assumed to be exponentially distributed as $\varsigma(z_1) = (1+a)e^{-(1+a)z_1}$ and if w is defined as

$$w = z_0 + z_1$$

then it is found in Aigner et al. [1977] and Meeusen and van den Broeck [1977] that

$$\begin{aligned} h(\ln u) = h(-w) = (1+a)\left[1 - \Phi\left(\frac{\sigma^2(1+a)+w}{\sigma}\right)\right] \\ \cdot \exp\left(\frac{\sigma^2(1+a)^2}{2} + (1+a)w\right) \end{aligned} \tag{4.27}$$

where $\Phi(\cdot)$ represents the cumulative distribution of the standard normal distribution.

If z_1 is assumed to have the half-normal distribution, we have[19]

$$h(\ln u) = h(-w) = \frac{2}{\sigma}\,\phi\left(\frac{w}{\sigma}\right)\left[1 - \Phi\left(\frac{w\lambda}{\sigma}\right)\right] \tag{4.28}$$

where $\phi(\cdot)$ is the standard normal density and

$$\sigma^2 = \sigma_{z_0}^2 + \sigma_{z_1}^2\,, \qquad \lambda = \sigma_{z_1}/\sigma_{z_0} \tag{4.29}$$

Inserting the various distributions $h(\ln u)$ in (4.10) yields the log-likelihood function of the sample.

According to Jondrow et al. [1982], conditional estimates of individual efficiencies can be obtained by computing expected values and modes

$$\begin{aligned} E(z_1/w) &= \frac{\sigma_{z_0}^2 \cdot \sigma_{z_1}^2}{\sigma^2}\left[\frac{\phi(w\lambda/\sigma)}{1-\phi(w\lambda/\sigma)} - \left(\frac{w\lambda}{\sigma}\right)\right] \\ M(z_1/w) &= \left(-w(\sigma_{z_1}^2/\sigma^2)\right) \quad \text{if } w \le 0 \\ M(z_1/w) &= 0 \quad \text{if } w > 0 \end{aligned}$$

[19] See Aigner et al. [1977].

in the normal case, and

$$E(z_1/w) = \sigma_{z_0}^2 \left[\frac{\phi(w/\sigma_{z_0} + \lambda^{-1})}{1 - \Phi(w/\sigma_{z_0} + \lambda^{-1})} - \left(\frac{w}{\sigma_{z_0}} + \frac{1}{\lambda} \right) \right]$$
$$M(z_1/w) = (-w - \sigma_{z_0}^2/\sigma_{z_1}) \quad \text{if } w \leq -\sigma_{z_0}^2/\sigma_{z_1}$$
$$M(z_1/w) = 0 \quad \text{if } w > -\sigma_{z_0}^2/\sigma_{z_1}$$

in the exponential case.

The choice between different functional specifications must, of course, be made on the basis of the information about the quality of the data, or on the basis of how the data are generated, and in accordance with the purpose of the study.

One word of caution seems in place with regard to the use of the composed error model. Consider a data set without measurement errors and where external shocks (weather, accidents, etc.) have not occurred. Now, if the specified efficiency distribution does not exactly mirror the observations, it is obvious that the symmetric (normal) distribution of the composed error structure will pick up some of the explanatory power of deviations from the frontier. Indeed, in the studies by Meeusen and van den Broeck [1977] and Aigner et al. [1977] the purely random term captures the lion's share of the variance of the composed stochastic variable. The point is that this may, to some extent, be accidental to the specified efficiency distributions.

4.5 Estimation via cost functions

Most applications of the frontier methodology have been to estimate production frontiers. However, estimation of production frontiers yields information on technical inefficiency but not on allocative inefficiency.[20] The behavioural assumption underlying direct estimation of the production frontier is generally the Zellner-Kmenta-Drèze assumption of expected profit maximisation, which implies exogenous input quantities.

It is well-known that the technology can be uniquely defined by either the profit function,[21] cost function or the production function. Which one is to be estimated depends on one's assumptions and/or data. The behavioural assumption underlying direct estimation of the cost function

[20] The estimation method uses data on input quantities but not input prices.

[21] An example of the use of a profit frontier can be found in Kumbhakar [1987].

is generally cost minimisation with exogenous output (e.g. because the firm is regulated). It requires data on input prices but not input quantities. Finally, the cost frontier yields information on the extra costs due to technical and allocative inefficiency, though not the separate cost of each without further assumptions.

To calculate allocative efficiency and overall efficiency, i.e., potential total reduction in average costs by moving to the frontier and adapting cost minimising factor ratios, factor prices are needed. But it does not necessarily follow that estimation via cost or factor demand functions is the only appropriate approach, as implied in Schmidt and Lovell (1979). The choice of approach must be determined by the economic mechanisms generating the data. For instance, if capital vintage effects are present, it may be impossible to infer the frontier cost function from the data if the relative factor prices have changed sufficiently since the date of investment. The data used in Schmidt and Lovell [1979] are for U.S. steam-electric generating plants, and the inputs are capital, fuel and labour. It seems unlikely that in the short run the capital variable in this type of plant can be as substitutable with the other factors as implied by the specified frontier cost function.

Cost frontiers can obviously be either deterministic or stochastic, just like production frontiers. Førsund and Jansen [1977] estimated a deterministic homothetic Cobb-Douglas cost frontier, with technical inefficiency represented by a density suggested earlier by Gabrielsen [1975].

A stochastic Cobb-Douglas cost frontier has been estimated by Schmidt and Lovell [1979]. The stochastic frontier model can also be extended so as to obtain separate estimates of technical and allocative inefficiency, provided that the functional form chosen for the production frontier is sufficiently tractable to permit derivation of the cost and input-demand frontiers in closed form. Schmidt and Lovell [1979] considered the Cobb-Douglas form of $f(\cdot)$ in (4.25)

$$\ln x = \ln A + \sum_{i=1}^{n} a_i \ln v_i - \omega - \theta \tag{4.30}$$

where the condition $\omega > 0$ allows for the occurance of production beneath the stochastic production frontier. In addition, they assume that the first-order conditions for cost minimisation are not satisfied. This is expressed by writing:

$$\ln(v_i/v_n) = \ln(a_i q_n / a_n q_i) + \varepsilon_i \qquad i = 1, \ldots, n-1 \tag{4.31}$$

where ε_i is symmetrically distributed, say multivariate normal with zero mean. If ε_i can be both positive and negative, then production is permitted to occur off the least cost expansion path. The combination of technical ($\omega \geq 0$) and allocative ($\varepsilon \gtreqless 0$) inefficiency yields a stochastic cost frontier of the form

$$\ln(q'v) = \beta_0 + \frac{1}{r}\ln x + \sum_{i=1}^{n}(a_i/r)\ln q_i + \frac{1}{r}(\omega + \theta) + E \tag{4.32}$$

where $r = \sum_{i=1}^{n} a_i$. Observed expenditure exceeds the stochastic cost frontier for two reasons: by an amount $\omega/r \geq 0$ due to technical inefficiency, and by an amount $E \geq 0$ due to allocative inefficiency.[22]

The model may be estimated by using MLE on the system of n equations in (4.30) and (4.31). The output of the estimation procedure consists of estimates of the frontier parameters $(\ln A, a_1, \ldots, a_n, r)$, the mean technical inefficiency over the sample $E(\omega)$, the extent of allocative inefficiency by observation, ε_i, the mean cost of technical inefficiency over the sample $(1/r)E(\omega)$ and the cost of allocative inefficiency by observation, E.

The model (4.30)–(4.32) can be extended in two directions. In the first place, the assumption that ε has a mean of zero can be replaced by the assumption that its mean is μ. This permits a test of the hypothesis that allocative inefficiency is systematic, $\mu \neq 0$ rather than random, $\mu = 0$. In the second place the assumption that technical and allocative inefficiency, ω and ε, are independent can be relaxed by permitting correlation between ω and $|\varepsilon|$. This permits a test of the hypothesis that firms which are relatively efficient technically are also relatively efficient allocatively.

The basic model (4.30)–(4.32) and both extensions are discussed in Schmidt and Lovell [1979, 1980], with an application to U.S. steam electric generation. The extended model is capable of shedding light on a wide variety of questions concerning the magnitudes and costs of technical and allocative inefficiency. It is, however, saddled with a fairly restrictive functional form, homogeneous Cobb-Douglas. In addition it should be pointed out that estimation of a system like (4.30)–(4.31) requires data on both input prices and input quantities, which may not always be available.

A system consisting of a deterministic translog cost frontier and the associated share equations has been estimated by Greene [1980b]. The advantage of the translog specification is its flexibility; a disadvantage is

[22] The term E is a well-specified function of the ε_i.

the impossibility of providing an explicit solution for the production function corresponding to the translog cost function or vice-versa. However, as shown in Kopp and Diewert [1982] and in Zieschang [1983] to decompose cost efficiencies into technical and allocative efficiency it suffices to know the frontier cost function. The basic relationship used is Shepard's lemma.

4.6 An example

As an example of estimating a deterministic frontier function with an explicit efficiency distribution, let us consider the cost function corresponding to the general homothetic function utilised in previous sections, and given in (4.4):

$$c = G(x)\Lambda(q_1, \ldots, q_n)u^{-1} \tag{4.33}$$

We shall consider here the exponential distribution corresponding to (4.11) and shown in (4.13) for the efficiency variable, u,

$$k(u) = (1 + a)u^a \tag{4.34}$$

Consider the case of cost minimisation with the reduced form (4.33). The distribution of efficiency with respect to costs is

$$d(u^{-1}) = d(z) = (1 + a)z^{-(a+2)}, \qquad z \in (\infty, 1] \tag{4.35}$$

This distribution is extremely simple: It contains only one parameter, but can still have shapes that are realistic enough for our purpose. The value of the parameter a determines the shape of the efficiency distribution: higher values of a imply that the expected observations are closer to the frontier.

The fundamental assumption of our model is that the firms have identical production functions, except for a term expressing technical efficiency in utilising the input index.

We are now interested in deriving maximum likelihood estimators for the parameters of the best-practice production function with distribution (4.34) as the efficiency distribution. According to the hypothesis of cost minimisation, we regard the output and the input prices as exogenous. The simultaneous distribution function of costs for a sample of N observations, assuming the random variables z^j, $j = 1, \ldots, N$, to be identically

and independently distributed with the distribution (4.35), is then

$$\ell(c^1,\ldots,c^N) = (1+a)^N \prod_{j=1}^{N} \left(G(x^j)\Lambda(q_1^j,\ldots,q_n^j)\right)^{(a+1)} \cdot (c^j)^{-(a+2)} \quad (4.36)$$

with

$$\infty > c^j \geq G(x^j)\Lambda(q_1^j,\ldots,q_n^j) \qquad j = 1,\ldots,N. \quad (4.37)$$

Maximum likelihood (ML) estimators for the parameters of the best-practice cost function are found by maximising $\ln \ell$ subject to constraint (4.37).

Since the maximising values of the parameters of the production function are independent of the value of the efficiency distribution parameter, the problem is to maximise:

$$\sum_{j=1}^{N} \ln\left(G(x^j)\Lambda(q_1^j,\ldots,q_n^j)\right) \quad (4.38)$$

subject to

$$\infty > \ln c^j \geq \ln\left(G(x^j)\Lambda(q_1^j,\ldots,q_n^j)\right) \qquad j = 1,\ldots,N. \quad (4.39)$$

(4.38) shows that when the basic cost function (4.33) in logarithmic form is linear in the parameters of both the transformation function and the price function, which is the case for the specification of $G(x)$ in (4.5) with Cobb-Douglas as the $g(v)$ kernel function in (4.2), then the maximum likelihood estimates of the parameters are obtained by solving a simple linear programming problem.

To describe the efficiency structure of the firms in the industry, we are interested in an estimate of the single parameter, a, characterising the efficiency distribution. Inserting ML estimates of the cost function parameters into the simultaneous distribution (4.36) in logarithmic form and differentiating with respect to a yields

$$\frac{N}{1+a} + \sum_{j=1}^{N} \left[\ln\left(\hat{G}(x^j)\hat{\Lambda}(q_1^j,\ldots,q_n^j)\right) - \ln c^j\right] = 0 \quad (4.40)$$

where the symbol $\hat{}$ indicates that ML-estimates are inserted. The ML-estimator for a is then

$$\hat{a} = \frac{1 - \frac{1}{N}\sum_{j=1}^{N}\left[\ln c^j - \ln\left(G(x^j)\hat{\Lambda}(q_1^j,\ldots,q_n^j)\right)\right]}{\frac{1}{N}\sum_{j=1}^{N}\left[\ln c^j - \ln\left(\hat{G}(x^j)\hat{\Lambda}(q_1^j,\ldots,q_n^j)\right)\right]} \quad (4.41)$$

The ML-estimator is a simple expression of the sum of deviations between observed costs and estimated best-practice costs. When (4.38) is solvable as a linear programming problem, these deviations are the value of each of the slack variables of the restrictions in (4.39).

4.7 Technical change and the frontier production function

Introduction

In this section we introduce technical change in a homothetic frontier production function by means of an illustrating example and based on the transformation function specified in (4.5) and a Cobb-Douglas kernel function. Then, we further develop Section 3.5, deriving measures of technical change for a homothetic frontier function.

It is shown in the case of a homothetic production function how the unit-cost reduction due to movement along a factor ray can be further split up multiplicatively into the reduction in unit cost due to the change in optimal scale, the cost reduction due to Hicks neutral technical change and the cost reduction due to factor bias technical change for a constant factor ratio.

Specification of technical change in the frontier production function

Consider the general time-dependent, homothetic function

$$G(x,t) = g(v,t) \cdot u \tag{4.42}$$

where x = rate of output, v = vector of inputs, $G(x,t)$ is a monotonically increasing function and $g(v,t)$ is homogeneous of degree 1 in v. The impact of technical change is simulated by assuming that the parameters of the functions in (4.42) are time functions. Using the same specification as in (4.6) and considering only two inputs for the sake of notational convenience, we have:

$$x^{\alpha-\gamma_4 t} e^{(\beta-\gamma_5 t)x} = A e^{\gamma_3 t} \prod_{i=1}^{2} v_i^{a_i-\gamma_i t} \cdot u \tag{4.43}$$

Technical change is accounted for by specifying the possibility of changes in the constant term A, and the kernel elasticities a_i, for v_i, and for the scale

function parameters α and β. The returns to scale properties are given by the scale elasticity function:

$$\varepsilon(x,t) = \frac{G(x,t)}{x \cdot G'(x,t)} = \frac{1}{\alpha - \gamma_4 t + (\beta - \gamma_5 t)x} \tag{4.44}$$

We assume that we have cross section time series data over T periods for N units, i.e., plants.

Let us consider the case of a deterministic frontier without any explicit specification of the efficiency distribution u. In accordance with the general purpose of frontier estimation, of fitting a frontier "as close as possible" to the observations, the computational model can be specified to minimise the simple sum of deviations from the frontier with respect to input utilisation after logarithmic transformation, subject to on or below frontier constraints. With this specification the estimation problem is reduced to solving a standard linear programming problem. The objective function to be minimised is:

$$\begin{aligned}\sum_{t=1}^{T}\sum_{j=1}^{N} \big(&\ln A + \gamma_3 t + (a_1 - \gamma_1 t)\ln v_1^j(t) + (a_2 - \gamma_2 t) \\ &\cdot \ln v_2^j(t) - (\alpha - \gamma_4 t)\ln x^j(t) - (\beta - \gamma_5 t)\cdot x^j(t)\big)\end{aligned} \tag{4.45}$$

where T is the number of periods and N the number of observations.

Note that although the objective function is linear in all the unknown parameters, the specification yields satisfactory flexibility as regards technical change.

Concerning the constraints of the LP-model, the expression within the brackets in (4.45) constitutes $(T \cdot N)$ constraints, securing the observed input points to be on or below the frontier:

$$\begin{aligned}&\ln A + \gamma_3 t + (a_1 - \gamma_1 t)\cdot \ln v_1^j(t) + (a_2 - \gamma_2 t)\cdot \ln v_2^j(t) \\ &\quad - (\alpha - \gamma_4 t)\cdot \ln x^j(t) - (\beta - \gamma_5 t)\cdot x^j(t) \geq 0\end{aligned} \tag{4.46}$$

In addition, we have the homogeneity constraint:

$$\sum_i a_{i,t} = \sum_i (a_i - \gamma_i \cdot t) = 1 \qquad t = 1, \ldots, T \tag{4.47}$$

Since (4.47) must be satisfied for all t, specification (4.43) implies the restriction:

$$\gamma_1 + \gamma_2 = 0 \tag{4.48}$$

It is not necessary to enter (4.47) for all T years because if it holds for one year, and (4.48) is valid, it must hold for all other values of t.[23] In addition we want the kernel elasticities including trends to be restricted to the interval $[0, 1]$. In view of (4.47) and (4.48) these constraints reduce to:

$$a_i - \gamma_i T' \geq 0 \qquad i = 1, 2 \tag{4.49}$$

We also want the scale parameters including trends to be non-negative

$$\alpha - \gamma_4 T' \geq 0 \tag{4.50}$$

$$\beta - \gamma_5 T' \geq 0 \tag{4.51}$$

We have found it reasonable in the empirical application of Chapter 7 to avoid the possibility of too abrupt a change in the scale function in the last year T, i.e., the optimal scale could be existent in the next to the last year but might not exist in the last year, by putting $T' = 2T$. Thus the non-negativity conditions will hold in the future for as long as the observed period lasts. This seems reasonable for prediction purposes.

Finally we have the following restrictions, which from an economic point of view seem reasonable:

$$\beta,\ \alpha,\ a_1,\ a_2,\ \gamma_3,\ \gamma_4,\ \gamma_5 \geq 0.$$

Note, however, that $\ln A$, γ_1 and γ_2 are unrestricted.

Technical change measures for a homothetic frontier function

For the homothetic function (4.2) the cost function is

$$c = G(x)\Lambda(q_1, \ldots, q_n) \tag{4.52}$$

where x is output and q_i, $i = 1, \ldots, n$, are the factor prices equal for both periods.[24] The technical advance measure (3.33)[25] then becomes:

$$T = G'_{x,t+1}(x^*_{t+1})\Lambda_{t+1}(q_1, \ldots, q_n) / G'_{x,t}(x^*_t)\Lambda_t(q_1, \ldots, q_n) \tag{4.53}$$

---

[23] Note that the choice of time index $t = 1, \ldots, T$ is not trivial. Our choice implies that the factor elasticities can never obtain extreme values for year 1 if the trends are different from zero.

[24] See, e.g., Førsund [1975].

[25] This measure is derived in Section 3.5.

Using the functional form in (4.43), optimal scale output x^* is:

$$x_t^* = (1 - \alpha_t)/\beta_t \tag{4.54}$$

Generally the conditional factor demand functions which correspond to the homothetic production functions are[26]:

$$v_i = \partial c/\partial q_i = G(x)\Lambda_i'(q_1, \ldots, q_n) \tag{4.55}$$

With a Cobb-Douglas kernel function as in (4.43) the calculation of the bias measure (3.37) becomes particularly simple. Using duality between production and cost functions yields the following expression for the price term $\Lambda(q)$ in (4.52):

$$\Lambda_t(q) = A_t^{-1} \prod_i (a_{i,t})^{-a_{i,t}} (q_i)^{a_{i,t}} \tag{4.56}$$

which yields Salter's bias measure:

$$D_{i,k} = \frac{\Lambda_{i,t+1}'(q_1, \ldots, q_n)/\Lambda_{k,t+1}'(q_1, \ldots, q_n)}{\Lambda_{i,t}'(q_1, \ldots, q_n)/\Lambda_{k,t}'(q_1, \ldots, q_n)} = \frac{a_{k,t}}{a_{k,t+1}} \cdot \frac{a_{i,t+1}}{a_{i,t}} \tag{4.57}$$

In Section 3.6 we also introduced the Binswanger bias measure, C_i, defined as the relative change in cost shares, for constant input prices and output level.[27]

This measure also becomes especially simple in our case, since due to (4.55) and (4.56) one obtains

$$C_i = \frac{v_{i,t+1}}{v_{i,t}} \cdot \frac{c_t}{c_{t+1}} = \frac{a_{i,t+1}}{a_{i,t}} \tag{4.58}$$

The Salter bias measure measures bias with respect to a certain factor. Thus, this measure is a function of the kernel elasticities of both factors under consideration. In contrast, the cost share measure of bias for a factor is simply given by the change in the kernel elasticity of this factor. This shows that these two ways of measuring bias might give different conclusions as regards the nature of the bias when more than two factors are involved.

In order to show the Farrell split-up of the unit-cost reduction into one part due to proportional shift towards the origin, and one part due to

[26] From Shepard's lemma.

[27] See (3.38).

the change in the optimal factor ratio, the factor ratios must be entered in (4.53) and (4.56) inserted. Consider the $n-1$ factor ratios

$$b_{ik} = v_i/v_k \qquad k = 1, \ldots, n \tag{4.59}$$

When these are given, all the other ratios follow. The prices generating them must then be:

$$q_k/q_i = a_k b_{ik}/a_i \qquad k = 1, \ldots, n \tag{4.60}$$

Substituting the price ratios in (4.53) and inserting (4.56), yields:

$$\begin{aligned} T = \frac{G'_{t+1}(x^*_{t+1})}{G'_t(x^*_t)} \cdot \frac{A^{-1}_{t+1}}{A^{-1}_t} \prod_k (D_{ki})^{-a_{k,t+1}} \\ \cdot \frac{a_{i,t}}{a_{i,t+1}} \prod_k (b_{ik})^{a_{k,t+1}-a_{k,t}} \end{aligned} \tag{4.61}$$

To find the proportional cost reduction part, T_1, we may calculate:

$$(v_{i,t+1}/x^*_{t+1})/(v_{i,t}/x^*_t) \tag{4.62}$$

We obtain $v_{i,t}$ and $v_{i,t+1}$ from (4.55) utilising (4.56) by inserting the factor ratios (4.59). These ratios are constant for t and $t+1$. When $\varepsilon_t(x^*_t) = 1$, we obtain $G'_t(x^*_t) = G_t(x^*_t)/x^*_t$ from (4.44). Using a Cobb-Douglas kernel function the result is

$$T_1 = \frac{G_{t+1}(x^*_{t+1})/x^*_{t+1}}{G_t(x^*_t)/x^*_t} \cdot \frac{A^{-1}_{t+1}}{A^{-1}_t} \cdot \prod_k (b_{ik})^{a_{k,t+1}-a_{k,t}} = OS \cdot H \cdot B \tag{4.63}$$

The first ratio, OS, shows the reduction in unit cost due to a change in optimal scale. The second term, H, shows the cost reduction due to the Hicks neutral technical change, and the third term, B, shows the cost reduction due to a factor bias technical change for a constant factor ratio.

In view of (4.57), the bias cost reduction part, T_2, must then be:

$$T_2 = \prod_k (D_{ki})^{-a_{k,t+1}} \cdot \frac{a_{i,t}}{a_{i,t+1}} \tag{4.64}$$

The Hicks factor neutral term, H, and the change in the scale function, OS, affect only the labelling of the isoquants, so they naturally belong to the proportional change term, T_1. Note that this term depends on the factor prices (factor ratios), but that the bias cost reduction term, T_2, is

independent of the factor prices. The latter term is, naturally, made up of a combination of the trends in the kernel elasticities.

The time functions specified in (4.43) are:

$$\begin{aligned} &a_1(t) = a_1 - \gamma_1 t\,, \qquad a_2(t) = a_2 - \gamma_2 t\,, \qquad \gamma_1 = \gamma_2\,, \\ &A(t) = Ae^{\gamma_3 t}\,, \qquad \alpha(t) = \alpha - \gamma_4 t\,, \qquad \beta(t) = \beta - \gamma_5 t \end{aligned} \tag{4.65}$$

With the two inputs utilised here, the technical advance measure (4.61) becomes

$$\begin{aligned} T = &\left(\frac{e\big(\beta - \gamma_5(t+1)\big)}{1 - \big(\alpha - \gamma_4(t+1)\big)}\right)^{1-(\alpha-\gamma_4(t+1))} \left(\frac{e(\beta - \gamma_5 t)}{1 - (\alpha - \gamma_4 t)}\right)^{(\alpha-\gamma_4 t)-1} \\ &\cdot e^{-\gamma_3} \cdot (b_{21})^{-\gamma_2} \cdot (D_{21})^{-a_2-\gamma_2(t+1)} \cdot \frac{a_1 - \gamma_1 t}{a_1 - \gamma_1(t+1)} \end{aligned} \tag{4.66}$$

The bias measures follow from inserting the time functions (4.65) into (4.57) and (4.58).

Remembering that $\sum_i \gamma_i = 0$ and $\sum_i a_{i,t} = 1$ for each t, in the case of a Cobb-Douglas kernel function with n inputs and time functions as specified in (4.65), B and T_2 may be expressed in the following way

$$B = \prod_k (b_{ik})^{-\gamma_k} = \prod_i v_i^{\gamma_i} \tag{4.67}$$

$$\begin{aligned} T_2 &= \prod_k (D_{ik})^{-(a_k - \gamma_k(t+1))} \cdot \frac{a_i - \gamma_i t}{a_i - \gamma_i(t+1)} \\ &= \prod_i \left(\frac{a_i - \gamma_i t}{a_i - \gamma_i(t+1)}\right)^{a_i - \gamma_i(t+1)} \end{aligned} \tag{4.68}$$

4.8 Concluding remarks

Consider the situation in which we have data on a cross-section of firms in an industry. The data includes output, and the prices and quantities of some inputs. In such a setting, it is natural to write a system consisting of the production (or cost) function, and of the first-order conditions for either profit maximisation or cost minimisation. None of these equations will fit the data perfectly, thus disturbances must be added. The question

is what does theory tell us about the nature and interpretation of these disturbances.

It is easier to talk about the disturbances in the first-order conditions. A standard assumption is that these are normal, that theory does not really dictate their form. They may be viewed as a measure of allocative inefficiency: if the technology is so simple that we can derive the explicit cost function from the production function and the first-order conditions, we see how much this raises cost. But the interesting question is, of course, relative to what state of the world cost is raised. If allocative inefficiency represents mistakes, cost is raised by these mistakes, and this is easy to interpret. But suppose, on the other hand, that we have a putty-clay situation in which the cost-minimising strategy of the firm dictates that a new plant will be built only occasionally, and in which, once the plant is built, certain input substitutions are impossible until the next plant is built. If relative input prices change, such a firm will be (some might prefer to say "appear to be") allocatively inefficient during any particular year of operation. But to say that this raises cost is wrong. Such a statement ignores the costs of adjustment which make the firm's strategy optimal.

It is not hard to find other similar examples. However, it should be noted that these examples do not argue against error terms on first-order conditions. They merely argue for caution in using the phrase "allocative inefficiency" to describe the phenomena they capture.

Next let us turn to the disturbance in the production function. Deterministic production frontiers are usually modelled with one-sided errors, while stochastic frontiers are modelled with two-sided errors. The one-sided error term represents production below the frontier, and is called technical inefficiency. Obviously it is possible to question this arrangement. Consider an idealised situation in which we observe every detail of the production process, including every conceivable input and every conceivable external circumstance (weather, strikes, disruption of supply, etc.). Then output would basically be deterministic. However, given a list of only, say, four inputs, output is certainly not exactly determined.The error term in the production function is an expression of this.

A deterministic or "pure" frontier uses a purely one-sided error. Hence it is assumed to be meaningful to be able to define exactly the maximal possible output, given some set of relevant inputs. Thus, for example, given quantities of seed, land, labour, fertilizer and capital, maximal output of corn for a farmer is assumed to be determined without error. Actual output is maximal output minus an inefficiency error. Clearly

this assumes that all other conceivable inputs or external events have a maximal possible (i.e., bounded) effect. For example, it is assumed that there is a best possible state of weather, a best possible set of farming practices, a best possible behaviour by insects, etc., so that under these best possible circumstances frontier output (but no more!) may be attained.

A stochastic frontier uses a mixture of one-sided and two-sided (e.g., normal) errors. Thus, given quantities of a list of inputs, there is a maximal output that is possible, but this maximal level is random rather than exact. This assumes that some other inputs or external effects have maximal possible effects, but others have potentially unbounded effects. For example, the effects of weather and other external events might be regarded as normally distributed (and thus unbounded). Thus the stochastic frontier expresses maximal output, given some set of inputs, as a distribution[28] rather than a point. However, it still must be possible to regard certain other inputs or external events as having maximal (best possible) values, so that their suboptimal values create the one-sided error.[29] Also, it should be stressed that statistical "noise" is found in every regression equation, and is usually argued to be normally distributed. This is just another reason for the stochastic nature of the frontier. For example, measurement errors on output fit in easily here, but create severe problems for a deterministic frontier.

Finally, it is possible to argue that there is no optimal value for everything, and hence there is no reason for a one-sided error or error component. In this view the concept of maximality is discarded, and a production function is regarded as merely giving the distribution of output, given certain inputs. If this view is accepted then there is no reason to study frontiers, of course.

Clearly those people who use frontiers must accept the notion of maximality. Since they often want to measure technical inefficiency, the failure to produce at the frontier is taken to be worth discovering, no matter what the reason for this failure.[30] This is true whether the frontier is deterministic or stochastic. Deterministic frontiers are often argued to be consistent with economic theory, but in fact their chief advantage seems clearly to be the availability of a measure of technical inefficiency for each observation.

[28] Typically, the distribution is normal.

[29] Typically these things would be those that are associated with the management practices of the firm.

[30] This is not to deny that, if possible, finding the reason for the failure would be worthwhile.

Their chief disadvantage is that they are bound to be confounded by statistical "noise". For stochastic frontiers the situation is exactly reversed. Thus, there is not yet a consensus on how one should, or whether one can, measure the technical efficiency of a firm, even if it is agreed that is a useful concept to measure.[31]

As with most philosophical discussions, this one may in the end be too pessimistic. Philosophical arguments have seldom prevented the use of techniques which yield plausible results. In that sense the real test of frontier models is likely to be an empirical one.

[31] For a further discussion, see Førsund [1985–86] and Schmidt [1985–86].

5

The Short-Run Industry Production Function

5.1 Introduction

The traditional assumptions in production theory of smooth (costless) substitution possibilities and costless choice of scale make it difficult to comprehend the structural development of several important industries more accurately characterised by quite limited substitution possibilities after the time of investment. The crucial difference between substitution possibilities before and after the actual construction of plants is most clearly captured by the vintage (putty-clay) approach, which assumes smooth substitution possibilities ex ante, and fixed coefficients for current inputs and capacity determined by the initial investment ex post.[1] The integration of these properties into a formal framework of production theory is found in Johansen [1972]. Within this framework it is necessary at the micro level, i.e., the unit of production, to distinguish between the production possibilities existing before the time of investment — the *ex ante production function* — and those existing after the investment — *the ex post production function.* Considering the short-run production possibilities for the entire industry as a unit, these must be based in some way on the individual *ex post* production functions. Aggregating, in a specific way as described below, the *ex post* functions of the micro units at a certain point in time yield *the short-run industry production function.*

The factors studied within the short-run function are limited to current inputs only. Fixed factors, such as capital, only determine the capacity of

[1] This basic model was analysed in Chapter 2.

the individual micro units and do not appear as variables in the short-run function. There are no costs associated with utilising the fixed factors in the short run.

The industry to which the short-run production function refers, comprises a certain number of production units N, $i = 1, \ldots, N$, characterised by a given output capacity. In the general case, which we shall look at first, *ex post* substitution possibilities between current inputs are allowed in the *ex post* micro functions

$$x^i = f^i(v^i, \bar{K}^i) \qquad i = 1, \ldots, N \tag{5.1}$$

where x denotes output and v a vector of actually used current inputs and $\bar{K}$ a vector of fixed factors.

The short-run industry production function is established by posing the classical problem of maximising output for given level of inputs. Thus, it corresponds to the basic definition of a production function when the industry is considered as one production unit, as opposed to the traditionally estimated function for an industry that was elaborated upon in Chapter 1. We are, let us remember, seeking a technical relationship independent of prices or economic behaviour. The short-run industry production function

$$X = F(V) = F(V_1, \ldots, V_n) \tag{5.2}$$

is obtained by solving the following problem:

$$\max_{x^i} X = \sum_{i=1}^{N} x^i \tag{5.3a}$$

subject to

$$x^i = f^i(v^i, \bar{K}^i) \qquad i = 1, \ldots, N \tag{5.3b}$$

$$\sum_{i=1}^{N} v_j^i \leq V_j \qquad j = 1, \ldots, n \tag{5.3c}$$

$$x^i \in [0, \bar{x}^i] \qquad i = 1, \ldots, N \tag{5.3d}$$

where X denotes output and $V_1, \ldots, V_n$ current inputs for the industry as a whole, $i = 1, \ldots, N$ refers to plants, and $\bar{x}^i$ denotes the capacity limit of unit i determined by $\bar{K}^i$. Free disposability of inputs is assumed.

Inserting (5.3b) into (5.3a) the necessary first-order condition of prob-

lem (5.3) is:

$$\frac{\partial f^i}{\partial v_j^i} - q_j - r^i \frac{\partial f^i}{\partial v_j^i} \leq 0 \qquad j = 1, \ldots, n \qquad i = 1, \ldots, N \qquad (5.4)$$

where q_j is the Lagrangean parameter associated with the constraint (5.3c) and r^i the one associated with (5.3d). For a fully utilised unit, (5.4) holds with equality. For a partly utilised unit (5.4) holds with equality when $r^i = 0$. Thus, the marginal productivities for a factor of partly utilised units are equal or less than for fully utilised ones. If (5.4) does not hold with equality for admissible input values the unit in question is not activated at all.

As regards the economic relevance of the derived short-run industry production function, as usual, two interpretations are possible:

(i) *Normative:* The short-run function shows how to organise the industry in the most efficient way when varying the degree of capacity utilisation and current factor prices, given that all units face the same factor and output prices.

(ii) *Positive:* The short-run function may simulate industry behaviour under decentralised decision making when all units face the same factor and output prices.

If it is not simulating actual market behaviour, the short-run function can, however, still be useful as a kind of description of industrial structure and structural change based on technical relationships, i.e., the distribution of input coefficients and capacity, giving a hypothetically maximum output for given amounts of inputs.

A series of short-run industry production functions over time are connected through the ex ante production functions. The ex ante function can be regarded as a choice-of-technique function for the construction of an *individual* micro unit. We can characterise it as a traditional production function with continuous substitution possibilities. Each production unit has at some time been "extracted" from the ex ante function that existed at that time. The short-run industry production function reflects both the history of ex ante functions over time and the actual choices made from these ex ante functions. Production at any point of time must be compatible with the short-run function.

5.2 Establishing the short-run industry production function

With respect to the assumptions about the ex post micro production functions, we employ here the same stylised vintage assumptions as in Chapter 2. Moreover, this seems also to be a reasonable approximation of the actual production possibilities of micro production units in several industries. The ex post micro functions are assumed to be of the following limitational type (where the unit index is deleted for simplicity):

$$x = \min \left[\frac{v_1}{\xi_1}, \ldots, \frac{v_n}{\xi_n}, \bar{x} \right] \tag{5.5}$$

where x is output, $\bar{x}$ the capacity limit, v_j the available amount of current input no. j, $\xi_j = \bar{v}_j/\bar{x}$, $j = 1, \ldots, n$, the input coefficient assumed to be constant and equal to the coefficient at full capacity utilisation and $\bar{v}_j$ the amount of input at full capacity.

In the following we assume that all micro units have the simple structure given by (5.5) but with different production capacities and different input coefficients. Empirically this seems to be a good approximation to reality. The input coefficients ξ_j are estimated by the observed coefficients.

The construction of the short-run industry production function can be formulated as

$$\max_{x^i} X = \sum_{i=1}^{N} x^i \tag{5.6a}$$

subject to

$$\sum_{i=1}^{N} \xi_j^i x^i \leq V_j \qquad j = 1, \ldots, n \tag{5.6b}$$

$$x^i \in [0, \bar{x}^i] \qquad i = 1, \ldots, N \tag{5.6c}$$

where X denotes output and $V_1, \ldots, V_n$ current inputs for the industry as a whole, and where $i = 1, \ldots, N$ refers to plants with a capacity of $\bar{x}^i$. Since for our purpose, we are only interested in the economic region, it has been natural to assume free disposability of inputs as expressed by (5.6b).

Another formulation of the short-run function is to proceed from the set of production activities describing ex post production functions by vectors of production activities at full capacity.[2] For a micro unit the produc-

[2] See Hildenbrand [1981] and Seierstad [1985].

tion activity at full capacity is

$$a = (\bar{v}_1, \ldots, \bar{v}_n, \bar{x}) \in R_+^{n+1} \tag{5.7}$$

The short-run production possibilities of the industry at a given point in time are then described by a finite family $\{a^i\}_{i \in N}$ of production activities. Given such a family of production activities we may define the short-run total production set

$$Y = (y \in R_+^{n+1} \mid y = \sum_{i \in N} \lambda^i a^i, \quad 0 \leq \lambda^i \leq 1) \tag{5.8}$$

where λ^i is the degree of capacity utilisation in unit i. A set of this type, i.e., a finite sum of line segments is called a *zonotope*.

Let D denote the projection of Y on the input space R_+^n, i.e.,

$$D = (V \in R_+^n \mid (V, X) \in Y \text{ for some } X \in R_+) \tag{5.9}$$

The short-run (efficient) industry production function $F : D \to R_+$ associated with Y is then defined by

$$F(V) = \max(X \in R_+ \mid (V, X) \in Y) \tag{5.10}$$

This formulation, however, does not allow free disposability of inputs without an easily undertaken redefinition of Y.

The optimisation problem raised above is a linear programming (LP) problem when the input coefficients are assumed constant. If they are functions of the capacity utilisation, which may be empirically relevant in some cases, a nonlinear programming problem arises.

In the sequel we will also need the dual problem of the LP-formulation (5.6). Let the dual variables $q_1, \ldots, q_n$ correspond to the restrictions on current inputs in (5.6b) and $r^1, \ldots, r^N$ correspond to the capacity limitations in (5.6c).[3] Let us first look at the following formulation of the whole problem:

The dual problem is to minimise

$$\sum_j q_j V_j + \sum_i r^i \bar{x}^i \tag{5.11}$$

[3] See Table 5.1.

Table 5.1: The linear programming tableau of the short-run function.

	Production units				Resources	Shadow prices
	1	2	...	N		
Coefficient matrix	ξ_1^1 ξ_2^1 $\vdots$ ξ_n^1	ξ_1^2 ξ_2^2 $\vdots$ ξ_n^2	... ... $\ddots$ ...	ξ_1^N ξ_2^N $\vdots$ ξ_n^N	V_1 V_2 $\vdots$ V_n	q_1 q_2 $\vdots$ q_n
	1 0 $\vdots$ 0	0 1 $\vdots$ 0	... ... $\ddots$ ...	0 0 $\vdots$ 1	$\bar{x}^1$ $\bar{x}^2$ $\vdots$ $\bar{x}^N$	r^1 r^2 $\vdots$ r^N
Activity levels	x^1	x^2	...	x^N		
Weights in the objective function	1	1	...	1		

subject to

$$\sum_j q_j \xi_j^i + r^i \geq 1 \qquad i, \ldots, N \tag{5.12}$$

The necessary first-order conditions are:

$$1 - \sum_{j=1}^{n} q_j \xi_j^i \gtreqless 0 \text{ when } \begin{cases} x^i = \bar{x}^i \\ x^i \in [0, \bar{x}^i] \\ x^i = 0 \end{cases} \qquad i = 1, \ldots, N \tag{5.13}$$

The variables, $q_1, \ldots, q_n$, are shadow prices of the current inputs with dimension output per unit of input. It follows directly that $q_1, \ldots, q_n$ represent the marginal productivities of the inputs of the industry function.

Whether a production unit is to be operated or not is then, according to (5.13), decided by whether current unit operation (dimensionless) "costs" calculated at these shadow prices are lower than or exceed unity. This corresponds to utilising units with non-negative quasi-rents. An equality sign in (5.13) defines the zero quasi-rent line, thus giving the boundary of utilisation of the set of production units in the input coefficient space. When operating costs equal unity, we have a marginal production unit in the sense that it may or may not be operated according to the optimal solution.[4]

5.3 Representation of the short-run industry production function

Introduction

Since the short-run production function is of a non-parametric form, the question of how the function should be represented now arises. This, of course, depends on the use to which the function is to be applied. In order to analyse long-run technical progress and structural change we need the complete representation of the substitution region, with each isoquant of the set found to be suitable for analysing three aspects: factor bias, productivity change and change in substitution properties.

Due to the linear structure of the problem (5.6a–c), the isoquants will be piecewise linear in the two-factor case considered here. In principle, the short-run function (5.6) can be derived numerically by solving a number of LP-problems. However, when the aim is to establish a reasonably interesting number of isoquants in order to reveal all the corners of the piecewise linear isoquants, solving the LP-problems (5.6a–c) is not a practical procedure.

If one is satisfied with the information given by a *limited* number of isoclines, these are readily obtained by utilising a simple ranking of the micro units according to unit production costs for given input prices. Such a cost minimisation procedure is utilised by Johansen [1972], K. Hildenbrand [1983] and W. Hildenbrand [1981]. The example of tankers given in Johansen [1972, Ch. 9] is an approximation which yields only a few of the corner points of the isoquant by a cost minimisation procedure based on four relative price ratios, while our recomputation on the same

[4] See Johansen [1972], pp. 13–19 for a more detailed exposition.

data revealed that one of the isoquants consisted of about two hundred line segments.

The algorithm for the construction of the isoquant

Our purpose is to establish a *complete* description of the substitution region and the isoquants. A complete description of the isoquants for the two-factor case is obtained by locating all the corner points geometrically, providing the whole set of isoclines and, thus in addition, enabling us to provide a full characterisation of the production function via marginal productivities, marginal rates of substitution, elasticities of substitution and elasticities of scale. Even for problems with a large number of production units the computation of isoquants is performed within a very reasonable amount of computer time. Briefly, the algorithm works as follows:

The boundaries of the substitution region are found by ranking the units according to increasing input coefficients for each input separately. This corresponds to ranking units according to unit costs when one input at a time has a zero price. An example is given in Figure 5.1, where also some isoquants are shown. The other characteristics of the figure are explained below. The industry in question is the Swedish cement industry in 1974, which is treated in detail in Chapter 8 and also utilised in Appendix 5.1. The complete data set, comprising 20 units, is shown in Figure 5.2 together with a transformation of the short-run function into the input-coefficient space. The current inputs are labour and energy.

The isoquants must be piecewise linear, downward sloping and convex to the origin, minimising costs for every factor price ratio. The essential idea is to substitute production units successively along the isoquant so that all these properties are fulfilled. This is obtained by the following geometric procedure:[5]

Starting from an arbitrarily chosen output level on the upper boundary, the last unit entered on the boundary is partially utilised. The problem is to find the next corner point on the isoquant. The algorithm then compares the slopes of the connecting lines in the input-coefficient space between the starting unit and all other units, i.e., all possible connecting lines between the units in Figure 5.2, and among the set of negative angles, picks out the unit yielding the steepest slope of the first isoquant line segment. Thus, two units are always partially utilised along an isoquant segment.

[5] Although addressed to other aspects of the short-run function, this geometric approach was inspired by an unpublished work by Seip [1974].

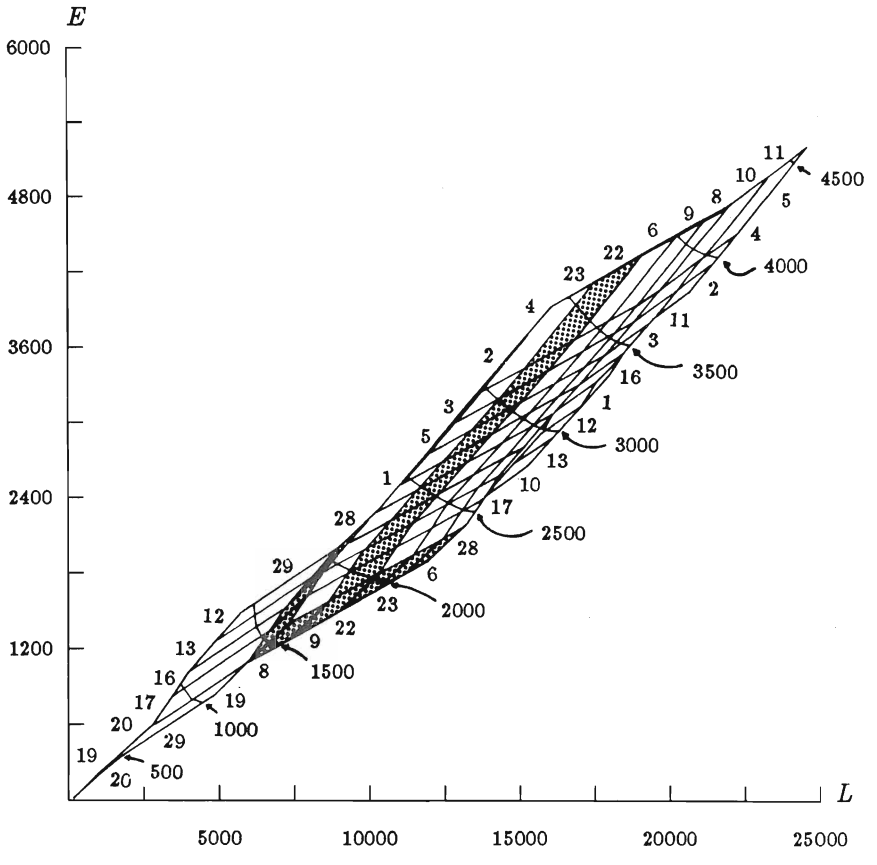
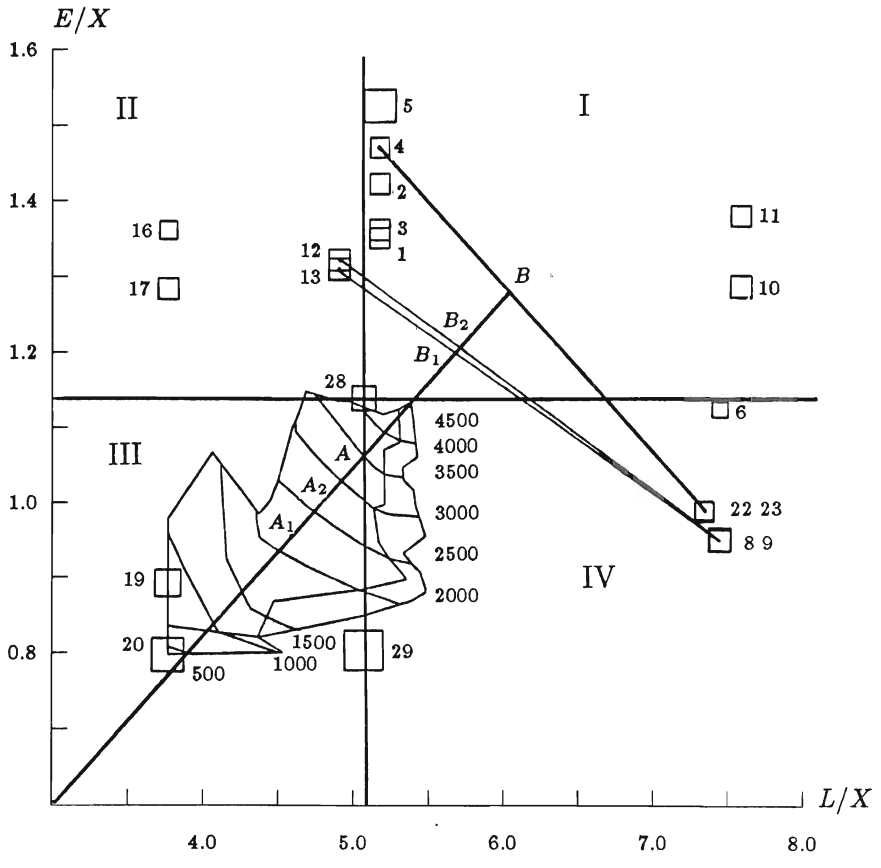


Figure 5.1: The short-run industry production function of the Swedish cement industry in 1974: the region of substitution, isoquants and activity regions.

In the case of increased utilisation of the starting unit, when moving from the boundary along the isoquant segment, the first isoquant corner point is reached either when capacity of the starting unit is exhausted, or when the capacity utilisation of the decreasing unit reaches zero. When the capacity



Note that the axes are truncated at the origin.

Figure 5.2: Capacity distribution and capacity region of the Swedish cement industry in 1974 with expansion path. Geometric calculation of the scale elasticity.

decreases, the corner point is reached when the utilisation of this unit reaches zero, or the utilisation of the increasing unit reaches 100 percent.

At each corner only one unit is partly utilised. The first segment can, at most, be vertical because the boundary units are sorted according to increasing input coefficients of that input which is increasing along the isoquant towards the lower boundary. The actual length of the segment depends on the capacity of the activated units.

The next step is to compare the angles to all other units in the input-coefficient space with the partly activated unit at the previously found corner point. The slope of the next line segment is then determined by the unit giving the second steepest slope *compared* to the slope of the previous line segment, and so on, until the lower boundary is reached.

The successive slopes of the connecting lines in the input-coefficient space between the units activated along the isoquant are the same as the slopes of the line segments in the input space. Intuitively this can be grasped by considering the shadow price interpretations of the dual variables q_1 and q_2. We see immediately that the marginal rate of substitution for the variables V_1 and V_2 is

$$\frac{\partial F/\partial V_1}{\partial F/\partial V_2} = \left(-\frac{dV_2}{dV_1}\right)_{dX=0} = \frac{q_1}{q_2} \tag{5.14}$$

The marginal rate of substitution function is discontinuous at the corner points.

Bearing in mind the shadow price interpretation of q_1 and q_2, discussed in connection with (5.13), it should be noted that the connecting line in the input-coefficient space between the two units utilised along an isoquant segment is also the zero quasi-rent line.

The isoquant obtained according to the algorithm above is convex and is as "close" to the origin as possible. According to the construction principle of the isoquant, an identical isoquant would be obtained if for the same output level total industry costs were minimised for successive price ratios equal to the marginal rate of substitutions as computed in (5.14). If the same factor prices apply to all units, it is obvious that the solution of the primal problem (5.6) implies cost minimisation for each output level. Therefore, it is evident that the isoquants must represent solutions of the LP-problem. A full description of the algorithm is presented in Appendix 5.1.

When transforming an isoquant to the input-coefficient space, the marginal rate of substitution is invariant. In general, the equation for a transformed isoquant is:

$$x = f(\xi_1 x, \ldots, \xi_n x) \tag{5.15}$$

where $f(\cdot)$ is a general production function. Thus, the marginal rate of substitution in the input-coefficient space corresponding to the point of evaluation in (5.14) is also q_1/q_2.

In Figure 5.2 the isoquant map of Figure 5.1 is transformed to the input-coefficient space, together with the boundaries of the substitution region. Such a transformed substitution region into the input-coefficient space is termed the *capacity region* in the sequel. As shown in Figure 5.2 the range of input coefficients that is possible to realise on the short-run industry function is considerably smaller than for the capacity distribution of the individual units, and may show a quite different form. Each point within the capacity region shows the average input coefficients of the units utilised to obtain the corresponding output level.

The activity regions

While the isoquants give information about how the units are utilised *across* the substitution region, one also might be interested in knowing how the units are utilised on the margin when moving *along* the substitution region. In addition to representing the short-run function by a limited number of isoquants, it may also be useful to reveal the *complete* set of efficient combinations of the micro units. An example is given in Figure 5.1.[6]

Starting at zero industry production and expanding to full capacity utilisation the activity regions are formed by adding micro units in accordance with the requirement that at each point in the substitution region, maximum industry output is obtained. Each parallelogram is formed by combining two units. Within the parallelogram the utilisation rate is between zero and 1. Each line segment of the parallelograms represents the locus of isoquant corners. Therefore, the activity regions' representation contains the complete set of all possible isoclines.

Such an activity region representation of the substitution region allows one to follow each individual unit's utilisation as a function of the industry's capacity utilisation. Southwest of the parallelogram the unit is not utilised, northwest of the parallelogram it is fully utilised. As seen in Figure 5.1, each unit is moved in parallel shifts in a strip-like fashion from one boundary of the substitution region to the other. Two examples of such strips are given by the shaded areas in Figure 5.1. Within each strip the units are partly

[6] A highly stylised example of such an activity region construction is given in Johansen [1972], Figure 2.1, p. 17. The generalisation, however, is not as simple as Johansen suggested on p. 18.

utilised, while obviously the utilisation rate for the unit in question is zero to the left and one to the right of the strip, corresponding to the utilisation rates at the boundaries of each parallelogram.

In the two-factor case the actual construction of the activity regions utilises the same slope matrix as for the construction of isoquants. Starting at the upper boundary of the substitution region, with the unit in use at the chosen point, the units which are to be combined with this one as one moves along the strip to the boundary are simply found by inspecting the slopes of the connecting lines between this unit and all others.

Referring to Figure 5.2, we can take unit no. 28 as an example. The strip for this unit is shaded in Figure 5.1. Placing unit no. 28 in the origin the units are divided into the four quadrants shown in Figure 5.2. The units in the first quadrant have higher input-coefficients than unit no. 28 for *both* inputs, whereas the units in quadrants II and IV have lower input coefficients for one of the inputs. The units in quadrant III have lower input coefficients for both factors, and thus will be fully utilised and never appear within the strip for the unit in question. Starting at the upper boundary of the substitution region, all units in the second and third quadrants are fully utilised, since the efficient utilisation along the upper boundary is in accordance with increasing input coefficients for only one factor, labour. When looking for a unit to be efficiently combined with our starting unit no. 28, it is obvious that this must be found among the units in the second or fourth quadrant, i.e., among units whose connecting lines with unit no. 28 have negative slope. Among these units, efficiency requires that the one with the steepest slope be picked out. If this unit is found in the second quadrant, it is already fully utilised, implying that the strip must move towards *lower* isoquant levels than the output level reached along the boundary before putting the starting unit into use. If the unit is found in the fourth quadrant, the strip moves towards higher isoquant levels, since units in the fourth quadrant are not yet put into use. It is the former case that appears here, unit no. 12. The next step is to locate the unit with the steepest slope among the remaining units, unit no. 13, and to continue (with units nos. 16, 17, 8, 9, 22, 23 and 6) until the set of units in the second and fourth quadrant is exhausted. The lower boundary is then reached.[7] In Figure 5.1 the complete set of partial utilisation strips are exhibited. The slope matrix behind the construction is found in Table A5.2. A summary description is offered below.

[7] A detailed exposition is found in Appendix 5.1.

The first units to be utilised are those that are on the convex hull of the units in the input-coefficient space. In Figures 5.1 and 5.2 on the 3500 ktonnes isoquant these units have the numbers 19 and 20. (The labelling of the units here is consistent with that of Appendix 5.1 and Chapter 8.) Technically these units together with nos. 29, 22 and 23 are dry cement kilns, while nos. 8 and 9 are semi-dry kilns, and the rest, wet kilns. Along the lower boundary of the substitution region the units are utilised according to increasing energy-input coefficients. Thus unit no. 20 is followed by unit no. 29 and then the units nos. 19, 8, 9, 22 and 23, which exhausts the dry and semi-dry technology. The first wet kiln is no. 6.

Along the upper boundary the units are utilised according to increasing labour-input coefficients. Here the two dry kilns, units nos. 19 and 20, are followed by four wet kilns, nos. 17, 16, 13 and 12, before a dry kiln, no. 29, enters. We have only one kiln, no. 20, which is so efficient (the most energy efficient unit and the second most labour efficient unit) that it only appears in one parallelogram, the one closest to the origin, and is from there fully utilised. Unit no. 19 appears in two parallelograms, no. 29 in five and no. 8 in fourteen. If we look at the isoquant levels, it turns out that unit no. 29 enters on isoquants from 500 ktonnes until 2000 ktonnes before it is fully utilised all the time. Unit no. 8 enters on every isoquant between 1500 and 4000 ktonnes. Unit no. 8 with semi-dry technology is fairly energy efficient but very labour consuming. Hence, this unit enters early when energy is relatively expensive, i.e., along the lower boundary, but disappears when labour becomes relatively expensive close to the upper boundary. Not until the industry is close to full capacity utilisation is this unit utilised all the time regardless of relative factor prices. Due to the relatively high labour-input coefficients of unit no. 22 it is interesting to note that this modern dry technology is not fully utilised until we reach about 90 percent capacity utilisation in the industry.

Two typical patterns of utilisation are illustrated by the shaded strips for units nos. 28 and 9 in Figure 5.1. In one case the utilisation depends to a large extent on the relative price, as for unit no. 9. The difference between the isoquant levels at the two boundaries is large for such units. In the other case the isoquant levels at the boundaries are about the same. But this is not sufficient for the utilisation pattern to be only scale dependent and relative price independent. The partial utilisation strip must also roughly follow the same isoquant level. The strip for unit no. 28 in Figure 5.1 is precisely an example of starting and ending at about the same isoquant, but in between it covers such a great range of isoquant levels that the utilisation pattern is also relative price dependent.

5.4 Further characterisation of the short-run function

Introduction

The short-run industry production function can be further characterised by its scale and substitution properties. Our computer program calculates both the values of marginal productivities and elasticity of scale as well as the marginal rate of substitution and "elasticity of substitution". Since the isoquants are piecewise linear, however, the utilisation of these measures is not without problems, as will be demonstrated below.

The elasticity of scale

From the classical theory of production we have the following well-known relationship:

$$\varepsilon X = \frac{\partial X}{\partial V_1} V_1 + \frac{\partial X}{\partial V_2} V_2 = q_1 V_1 + q_2 V_2 \tag{5.16}$$

The first equation is the *passus equation* in the terminology of Frisch [1965], with ε the elasticity-of-scale function, which is discontinuous at all corner points. The second equation follows directly from the shadow price interpretation of the variables q_1 and q_2

Therefore, to be able to calculate the scale elasticity it is necessary to find q_1 and q_2. This is done by utilising the fact that (5.13) holds with an equality sign for marginal units. In the two-factor case there must be two marginal units on every isoquant segment; the utilisation rate of one is increasing and that of the other decreasing. On each segment we then have two equations, (5.13), in the two unknowns, q_1 and q_2.

Obviously the scale elasticity is constant along an isoquant segment:

$$d\varepsilon = 1/X(q_1 dV_1 + q_2 dV_2) = 0 \tag{5.17}$$

remembering that q_1 and q_2 are interpreted as marginal productivities and that (5.14), the property of a constant marginal rate of substitution along an isoquant segment, holds.

In the case of a continuous capacity distribution it is shown by Johansen [1972] that the scale elasticity may be given a geometrical interpretation in the input-coefficient space.[8] A similar geometric interpretation carries over to our case of discrete capacity distribution. Consider

[8] See Figure 4.2 in Johansen [1972].

again Figure 5.2. As pointed out in Section 5.3, each point on the transformed isoquants within the capacity region represents the average input coefficients, or the centre of gravity of the utilised "capacity mass" for that point on the short-run industry production function. Consider the point A on the 3500 ktonnes isoquant in Figure 5.2. Two micro units are partially utilised at this point and constitute the marginal units. The units are marked as 4 and 23 in Figure 5.2. Now, in the continuous capacity distribution case Johansen [1972] shows that the scale elasticity is found by taking the ratio between the average input coefficient and the coordinate point in the input-coefficient space, determined by the intersection of the ray from the origin through the average point with the zero quasi-rent line. In the discrete case the zero quasi-rent line, used in establishing the short-run function, will always pass through two units. The point of intersection, however, may be between the marginal units as well as outside both. If one wants to use the interpretation suggested by Johansen[9] of the scale elasticity as "the proportion between the input coefficients of the average production unit and the input coefficients of the marginal production unit with the same factor proportion", one should bear in mind that such a marginal unit cannot be obtained physically by combining the two real marginal units, since the weights in the linear combination along the zero quasi-rent line are not restricted to the $[0, 1]$ domain.

In Figure 5.2 the geometrical computation of the scale elasticity is illustrated. Corresponding to a chosen average point A, the marginal units are nos. 4 and 23, and the intersection point between the ray through A and the zero quasi-rent line through units nos. 4 and 23 is denoted by B. The value of elasticity of scale is obtained as $OA/OB = 0.84$. Note that the axes are truncated at the origin.

If we consider the average point at the upper boundary on the second isoquant segment of the same isoquant that point A is on (3500 ktonnes), it turns out that the marginal units are nos. 22 and 8.[10] The intersection point is in this case far to the left of these units. The fact that the scale elasticity is constant along an isoquant segment is confirmed here by geometry. When moving point A along an isoquant segment the marginal units remain the same, and the intersection point, B, moves along the zero quasi-rent line parallel to the isoquant segment in question.

As pointed out by Johansen,[11] the scale elasticity also shows the dis-

[9] See Johansen [1972], Section 4.4, p. 66.

[10] See line no. 2 in Table A5.3.

[11] See Johansen [1972], Section 4.4.

tribution of total value of production when evaluated at the shadow prices, q_i in (5.13). The share of quasi-rent of total value of production is equal to $(1 - \varepsilon)$ measured geometrically as AB/OB.

Regarding the variation of the scale elasticity, ε, the case of discrete distributions enables us to be more precise than in the continuous case formulated as follows by Johansen:

> Apart from (4.38) [$\varepsilon = OA/OB \to 1$ as $X \to 0$] and the fact that $\varepsilon < 1$ for $X > 0$, not much can be said in general about the variation in ε.
>
> In particular ε does not necessarily decrease monotonically with increasing output. It is easy to conceive of distributions — which are such that ε first decreases but later on passes through both increasing and decreasing phases as output X increases, although this may perhaps not be very realistic in practice.[12]

Considering now the regions of constant isoquant segment slopes in Figure 5.1 we have that for each such parallelogram the scale elasticity attains its smallest value as a factor ray enters from the origin and then *increases* within the region, attaining its largest value as the factor ray leaves the region. This is because the two marginal units are the same for the whole region, i.e., point A in Figure 5.2 moves outwards while point B is constant. Even if we compare two factor points in different parallelograms, the scale elasticity may also increase when moving outwards, depending upon the relative change of the average point A and the marginal point B. For example, this is the case in Figure 5.2 when factor points are compared along the average factor ray corresponding to the average input coefficient points A_1 and A_2 with marginal input coefficient points B_1 and B_2, respectively.

If "in practice" in the quotation above is interpreted as referring to the case of discrete capacity distributions, then, when output is increasing, both increasing and decreasing phases of the scale elasticity are certainly the rule.

The curvature of the isoquants

How should the curvature of the isoquants be characterised in this case with piecewise linear isoquant segments? Considering the very purpose of studying the curvature of the isoquants, interesting questions are:

---

[12] Johansen [1972], p. 67.

(a) How large is the saving of one input, say energy, when moving along different parts of the isoquants?
(b) How sensitive are the quantities of inputs to a change in relative input prices along the isoquant?
(c) Is it possible to find parts of the isoquant where small changes in relative prices yield large changes in input quantities or vice versa?

Since the isoquants consist of piecewise linear segments it is difficult to find numerical measures confirming the visual impression of the curvature of an isoquant.

The conventional measure of substitution properties, the elasticity of substitution, is zero at the corner points and infinity along the segments. One possibility is to approximate the isoquant with a smooth curve and compare this form of the isoquant with isoquants of a well-known analytical production function. This is performed in Førsund and Hjalmarsson [1978b]. Another possibility is, by analogy with the definition in the case of smooth isoquants, to compute the elasticity of substitution as an arc elasticity for two consecutive isoquant segments in the following way:

$$\sigma^s = \frac{\dfrac{V_2^s}{V_1^s} - \dfrac{V_2^{s+2}}{V_1^{s+2}}}{\dfrac{V_2^s}{V_1^s} + \dfrac{V_2^{s+2}}{V_1^{s+2}}} \cdot \frac{\dfrac{1}{2}\left(\dfrac{V_2^s - V_2^{s+1}}{V_1^s - V_1^s} + \dfrac{V_2^{s+1} - V_2^{s+2}}{V_1^{s+2} - V_1^{s+1}}\right)}{\dfrac{V_2^s - V_2^{s+1}}{V_1^{s+1} - V_1^s} - \dfrac{V_2^{s+1} - V_2^{s+2}}{V_1^{s+2} - V_1^{s+1}}} \qquad (5.18)$$

$$s = 1, \ldots, S-2$$

where V_1^s and V_2^s are the coordinate values at corner point no. s, where S is the number of corner points along the isoquant. Thus, we have utilised the average factor ratio and average slopes for pairwise isoquant segments yielding $S-2$ values of the elasticity of substitution.

However, when the number of isoquant segments is high, the number of elasticities of substitution for the same isoquant is also high. Moreover, as the empirical results below illustrate, the value of the elasticity of substitution varies considerably and unsystematically among pairwise isoquant segments. An alternative would be to take an arc elasticity, according to the same principle as in (5.18), but for more than two segments at the same time. One should still, however, expect the elasticity of substitution to vary considerably along the isoquant.

A fourth possibility is, of course, to look directly at the slope of the isoquant segments, q_1/q_2, i.e., to look at the marginal rate of substitution.

There does not seem to be any easy way of summarising all the substitution properties of the whole isoquant since the arc elasticity is as detailed as the isoquant itself. If very detailed information about the substitution properties of limited parts of the isoquant is needed, the arc elasticity of substitution serves quite well. If we are interested in summary information, there is no obvious way of either fitting a smooth isoquant or parametricising an elasticity of substitution function.

The end-points of the isoquants give us the scope for factor substitution, and answer questions like how much is the maximal possible reduction in one input, say energy, for constant output. In a short-term policy context the substitution possibilities between the inputs for a given level of output can be of great interest, for example, in an energy crisis or when analysing industrial policy problems.

Appendix 5.1: The isoquant plotting algorithm

The algorithm can be described by the following steps:

1. Data requirements for each unit:

   current output: x
   capacity: $\bar{x}$
   current inputs: v_j, $j = 1, 2$.

2. Calculate all input coefficients and sort them in increasing order of ξ_1 (an arbitrary choice) and renumber according to this sorting.

3. Calculate all slopes of the connecting lines between the micro units in the input-coefficient space. The slopes are denoted by $S_{k\ell}$

$$S_{k\ell} = \frac{\xi_{2k} - \xi_{2\ell}}{\xi_{1k} - \xi_{1\ell}} \qquad \begin{matrix} k = 1, \ldots, N-1 \\ \ell = k+1 \ldots, N \end{matrix}$$

   In the case of the unlikely event of zero in the denominator, either the abscissa and ordinate variables have to be changed, or the denominator must be given an arbitrarily small increment.
   The S-coefficients are gathered in a triangular matrix without the main diagonal where the units are entered according to increasing input coefficients, ξ_1, of the abscissa input variable both along the rows and columns.

Table A5.1: The slope-matrix.

k \ ℓ	2	3	.....	N
1	S_{12}	S_{13}	.....	S_{1N}
2		S_{23}	.....	
3			$\ddots$	
$\vdots$			$S_{k\ell}$	
$\vdots$			$\ddots$	
$N-1$				$S_{N-1,N}$

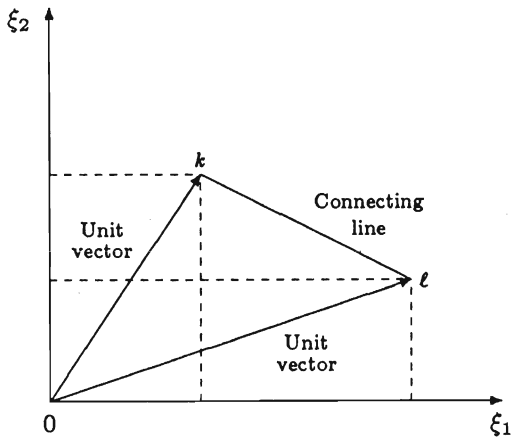


Figure A5.1: Calculation of slopes between units in the input-coefficient space.

4. Choose isoquant level, X^0. Isoquant plotting starts from the upper boundary (an arbitrary choice). The unit partly in use is identified by finding the smallest i that satisfies $\sum_i \xi_{1i} \cdot \bar{x}_i \geq X^0$.

5. Starting now from a chosen output level on the upper boundary, the

last unit entered on the boundary is partially utilised. The problem is to find the next corner point on the isoquant. The algorithm then compares the slopes of the connecting lines between the starting unit and all units in the input-coefficient space.

Referring to Table A5.1, this means that the algorithm inspects the figures in the column for the starting unit (e.g., column ℓ) and the figures on the row $\ell + 1$ for the same unit. Thus the algorithm picks out the unit in the table yielding the steepest slope of the first isoquant segment by locating the largest absolute value of the negative slopes in the column ℓ and the row $\ell + 1$.

The column ℓ contains all utilised units, whereas the row $\ell+1$ consists of units which are not utilised. If the largest figure is found in the column, e.g., $S_{k\ell}$, the capacity utilisation of the starting unit found in column ℓ is increased. At the same time, the capacity utilisation of the unit on the row k is decreased. If the largest figure is found in the row $\ell+1$, the capacity utilisation of the starting unit is reduced while the capacity utilisation of the unit in the corresponding column (e.g., $\ell + t$ where $t \in [\ell, N - 1 - \ell]$) is increased.

In the case of increased utilisation of the starting unit, the first isoquant corner point is reached when either the capacity of the starting unit is exhausted or the capacity utilisation of the decreasing unit reaches zero. When the capacity utilisation of the starting unit decreases the corner point is reached when the utilisation of this unit reaches zero or the utilisation of the increasing unit reaches 100%. At the corner only one unit is partly utilised. The first segment can at most be vertical because the boundary units are sorted according to increasing ξ_i-input coefficients of that input which is increasing along the isoquant towards the lower boundary. The actual length of the segment depends on the capacity of the activated units.

The next step is to compare the angles of all other units in the input-coefficient space with the partly activated unit at the previously found corner point. The angle of the next line segment is then determined by the unit giving the steepest angle next to the angle of the previous line segment. The process is repeated until the lower boundary is reached.

Illustrating example

The purpose of this section is, by means of a numerical example, to give a more detailed presentation of the algorithm for the computation of the short-run industry production function. The example refers to the

3500 ktonnes isoquant of the cement industry in 1974, illustrated in Figure 5.1. The complete slope matrix is presented in Table A5.2. The boundary of the substitution region up to this isoquant level and the isoquant itself is presented in Table A5.3.

The Substitution Region

The boundaries of the substitution region are found by ranking the units according to increasing input coefficients for each input separately. This corresponds to sweeping horizontal and vertical "price" lines outwards from the axes over the capacity distribution, as seen in Figure 5.2, and entering the plant capacities in the order they appear. In Table A5.3 the units are ranked according to increasing labour input coefficients on the upper boundary. On the first 14 rows in Table A5.3, the upper boundary of the substitution region is built up for both the increments in labour L and energy E, and the accumulated values are printed out. The last unit entered at the starting point of the isoquant, Unit no. 23, is utilised to 54.7% of its capacity.

Isoquants

Starting now from the chosen output level on the upper boundary, the last unit entered on the boundary is partially utilised. In Table A5.3 this unit is no. 23. In Table A5.2, the algorithm inspects the quantities in the column for the starting unit, no. 23, and the quantities on the row for the same unit. For convenience absolute values are used in this discussion. Thus the algorithm picks out the unit in the table yielding the steepest slope of the first isoquant segment by locating the largest quantity either in the column or the row for the starting unit, no. 23. In this example the quantity is 315.81 in the row for unit no. 23. This quantity also appears in the column for unit no. 22. Since the largest quantity is found in the row, the capacity utilisation of the starting unit no. 23 is decreased from 54.7% to zero.

At the same time, the capacity utilisation of the unit no. 22 is increased from zero to 48.3%. The first corner point of the isoquant is reached when the capacity of the contracting unit no. 23 is zero.

In our example the next line segment and corner point is found by inspecting the quantities in the column of unit no. 22 and the row for the same unit. The largest figure, not exceeding 315.81, is 0.36 in the column of unit no. 8. Since the largest figure is found in the row of unit no. 22, the capacity utilisation of unit no. 22 is decreased from 48.3% to zero. At

Table A5.2: The S-matric of slope coefficients for the Swedish cement industry in 1974.

	20	17	16	13	12	29	28	1	5	3	...
19	−1054.24	2756.04	1804.77	0.36	0.37	−0.07	0.19	0.32	0.44	0.33	
20		9298.85	3313.17	0.44	0.45	0.01	0.26	0.39	0.51	0.40	
17			657.10	0.02	0.03	−0.36	−0.11	0.05	0.17	0.06	
16				−0.04	−0.03	−0.42	−0.17	−0.01	0.12	0.00	
13					394.90	−3.07	−1.04	0.19	0.80	0.19	
12						−3.14	−1.11	0.11	0.75	0.15	
29							9765.03	5.13	6.77	5.23	
28								1.99	3.63	2.09	
1									5945.69	252.50	
5										−11542.25	
⋮											

...	2	4	23	22	6	9	8	10	11	
	0.37	0.41	0.03	0.03	0.06	0.02	0.02	0.10	0.13	19
	0.44	0.47	0.05	0.05	0.09	0.04	0.04	0.13	0.15	20
	0.10	0.13	−0.08	−0.08	−0.04	−0.09	−0.09	0.00	0.03	17
	0.04	0.08	−0.10	−0.10	−0.06	−0.11	−0.11	−0.03	0.01	16
	0.41	0.59	−0.13	−0.13	−0.07	−0.14	−0.14	−0.03	0.03	13
	0.37	0.55	−0.13	−0.14	−0.08	−0.14	−0.15	−0.11	0.02	12
	5.79	6.23	0.08	0.08	0.13	0.06	0.06	0.19	0.23	29
	2.65	3.10	−0.06	−0.07	−0.01	−0.08	−0.08	0.06	0.10	28
	390.55	428.36	−0.17	−0.17	−0.10	−0.17	−0.18	−0.01	0.01	1
	−675.32	−221.43	−0.25	−0.25	−0.18	−0.25	−0.25	−0.10	−0.06	5
	433.83	460.93	−0.17	−0.17	−0.10	−0.18	−0.18	−0.03	0.01	3
		500.00	−0.20	−0.20	−0.13	−0.21	−0.21	−0.06	−0.02	2
			−0.22	−0.22	−0.15	−0.23	−0.23	−0.08	−0.04	4
				−315.81	1.26	−0.36	−0.38	1.20	1.58	23
					1.28	−0.34	−0.36	1.21	1.59	22
						−95611.04	−1042.20	1.16	1.82	6
							−12.04	2.38	3.05	9
								2.40	3.07	8
									539.47	10

Table A5.3: The construction of an isoquant: 3500 ktonnes in 1974.

Line no	Unit in	Type	Fraction Before	Fraction After	Unit out	Fraction Before	Fraction After	Increments in E	Increments in L	Slope	Coord. Sum E	Values Sum L	Comment on corner
	19	dry	zero	one	none			263.0	1111.6		263.0	1111.6	contour
	20	dry	zero	one	none			355.4	1679.1		618.4	2790.8	contour
	17	wet	zero	one	none			224.9	662.3		843.4	3453.1	contour
	16	wet	zero	one	none			184.9	513.7		1028.3	3966.8	contour
	13	wet	zero	one	none			244.3	919.1		1272.6	4885.9	contour
	12	wet	zero	one	none			241.6	900.8		1514.2	5786.7	contour
	29	dry	zero	one	none			512.7	3237.1		2026.9	9023.8	contour
	28	wet	zero	one	none			279.6	1249.4		2306.5	10273.2	contour
	01	wet	zero	one	none			233.9	900.2		2540.5	11173.5	contour
	05	wet	zero	one	none			665.3	2233.7		3195.7	13407.1	contour
	03	wet	zero	one	none			242.2	924.4		3437.9	14331.6	contour
	02	wet	zero	one	none			246.2	900.3		3684.2	15231.8	contour
	04	wet	zero	one	none			243.6	861.6		3927.8	16093.5	contour
	23	dry	zero	0.547	none			77.0	571.6		4004.8	16665.1	isoq. start
1	22	dry	zero	0.483	23	0.547	zero	−0.2	0.0	315.813	4004.6	16665.1	isoq.
2	08	semi-dry	zero	0.417	22	0.483	zero	−2.9	8.2	0.360	4001.7	16673.3	isoq.
3	08	semi-dry	0.417	one	05	one	0.748	−62.0	246.4	0.252	3939.7	16919.7	isoq.
4	09	semi-dry	zero	one	05	0.748	0.246	−123.0	490.5	0.251	3816.7	17410.2	isoq.
5	22	dry	zero	0.659	05	0.246	zero	−56.5	229.5	0.246	3760.2	17639.7	isoq.
6	22	dry	0.659	one	04	one	0.670	−26.2	118.7	0.221	3734.0	17758.4	isoq.
7	23	dry	zero	0.783	04	0.670	zero	−53.0	240.9	0.220	3681.0	17999.3	isoq.
8	23	dry	0.783	one	02	one	0.822	−13.2	67.0	0.198	3667.7	18066.2	isoq.
9	06	wet	zero	one	02	0.822	0.233	−30.2	231.9	0.130	3637.5	18298.1	isoq.
10	10	wet	zero	0.222	02	0.233	zero	−5.40	97.5	0.056	3632.1	18395.6	isoq.
11	10	wet	0.222	one	03	one	0.203	−10.5	342.3	0.031	3621.6	18737.9	isoq. end

the same time the capacity utilisation of unit no. 8 is increased from zero to 41.7%.

In the third step the quantities in the column and row of unit no. 8 are inspected. It turns out that the largest figure not exceeding 0.36 is 0.25, found in the column of unit no. 8 and on the row of unit no. 5. This means that the capacity utilisation of unit no. 8 is increased to 100% and that of unit no. 5 is decreased.

Since unit no. 5 is still partly utilised, in the fourth step the column and row of unit no. 5 is inspected again. It turns out that unit no. 9 is the next unit involved. And so the process repeats itself until no further point can be found. Then, the last point is found on the lower boundary. "Limiting" cases are covered by special routines.

Note that sometimes along an isoquant a unit may drop out totally to return later on the same isoquant. This holds for both unit no. 22 and unit no. 23. It is only possible to identify nine of the eleven isoquant segments in Figure 5.1 due to the almost equal labour coefficients of units nos. 8, 22 and 23, which lead to very narrow parallelograms.

The activity regions

The location of the activity regions, as in Figure 5.1, follows from a straightforward utilisation of the slope matrix. The substitution region may be filled up with activity regions by entering in turn strips of parallelograms for each micro unit. Choosing an arbitrary unit, the units to be combined with it are found by inspecting the corresponding column, for example, ℓ in Table A5.1 and row $\ell + 1$ for negative slopes. The units corresponding to the slopes in a column are found in the second quadrant in Figure 5.2 and the slopes in a row correspond to units in the fourth quadrant.

The first unit to be combined with the chosen unit is the one with the largest absolute slope value. Then the other units are combined in descending order of the slope values. When a slope value is picked from the column, the corresponding parallelogram is formed by *subtracting* the full capacity input values of the unit in question from the previously obtained coordinate values in the substitution region, respectively representing zero and full capacity utilisation of the chosen unit. When a slope value is picked from the row, the parallelogram is obtained by *adding* the full capacity input values. Thus, a partial utilisation strip changes direction each time the picking of consecutively decreasing slope values changes from row to column, or vice versa.

As an example, let us consider unit no. 28. The connecting line with the largest absolute slope, 1.11, is with unit no. 12 (as can be seen in Figure 5.2) and found in the column of no. 28 in Table A5.2.

The strip therefore starts in the direction of the origin. The next three slopes are also found in the column, but then we jump to the row and stay there, continuing with unit no. 8 in the direction *from* the origin until all units with negative slopes have been combined with unit no. 28.

The properties of the production function

The computer program provides very detailed information about the isoquants. As an illustrating example let us again look at the isoquant presented in Table A5.3, representing the capacity level of 3500 ktonnes cement in 1974. In Table A5.4 the values of the elasticity of scale and the elasticity of substitution are listed along with the marginal productivities of labour and energy, the marginal rate of substitution and the factor ratio for each line segment of the isoquant, which are sorted from the upper boundary to the lower boundary. The total isoquant consists of 11 line segments.[13]

In the columns of the marginal productivities of labour and energy we find that the productivity of labour increases from about 0.01 tonnes per hour on the upper boundary line to about 6.6 tonnes per hour on the lower boundary, at the same time as the marginal productivity of energy decreases from 0.14 tonnes per Gcal. to about 0.02 moving along the isoquant in the labour intensive direction. The actual factor price ratio this year is 0.11, tangential to an isoquant corner close to the lower boundary.

Looking at the elasticity of scale column, we see that the scale elasticity varies somewhat along the isoquant, increasing and then decreasing from the upper boundary. This seems to be a typical pattern confirmed in Chapter 8.[14]

The information of the marginal rate of substitution or factor price ratio column and the factor ratio column is combined in the elasticity of substitution column. The value of the elasticity of substitution varies considerably and unsystematically along the isoquant. Hildenbrand [1981]

[13] It might be mentioned that this is an example of a *product table* in the terminology of Frisch [1965], Ch. 5.a, i.e., a numerical representation of the production function.

[14] The value of the scale elasticity on the seventh isoquant segment, 0.84, was also calculated above in Figure 5.2.

Table A5.4: A characterisation of the 3500 ktonnes isoquant of the Swedish cement industry in 1974.

Isoquant line segment no.	Factor price ratio energy/ labour	Marginal productivity of energy ktonnes/ Tcal.	Marginal productivity of labour ktonnes/ 100 hrs.	Factor ratio Tcal./ 100 hrs.	Elasticity of scale	Elasticity of substitution
1	315.8130	0.1357	0.0009	0.240	0.6471	0.0003
2	0.3604	0.0989	0.2744	0.240	0.7850	0.0443
3	0.2516	0.0889	0.3532	0.240	0.8272	13.3601
4	0.2507	0.0887	0.3538	0.233	0.8270	2.4691
5	0.2463	0.0878	0.3567	0.219	0.8260	0.1922
6	0.2209	0.0844	0.3822	0.213	0.8361	4.9912
7	0.2200	0.0842	0.3829	0.210	0.8358	0.1655
8	0.1979	0.0808	0.4083	0.205	0.8450	0.0343
9	0.1302	0.0621	0.4768	0.203	0.8201	0.0173
10	0.0555	0.0325	0.5849	0.199	0.7778	0.0243
11	0.0307	0.0202	0.6578	0.197	0.7886	

claims that as a "general empirical fact" (his quotation marks) the values of this elasticity are quite low. However, although there are many very low values in Table A5.4, the values vary considerably up to quite high values, and it is difficult to read off any systematic pattern. For two pairwise segments the elasticity is rather high around 13.36 and 4.99, respectively, but for other segments it is extremely low, below 0.01. Comparing these results with the graph of the isoquant in Figure 5.1 illustrates the difficulties of using these elasticity figures as a summary description of the curvature of the whole isoquant.

The 3500 ktonnes isoquant in Figure 5.1 is somewhat L-shaped. From Table A5.3 it is easy to calculate that along the first seven line segments it is possible to decrease energy input by 8.1% by increasing labour consumption 8.0%. On the other hand, on the last four segments it is possible to decrease energy input by 1.3% by increasing labour consumption 3.7%.

Moving from the lower end point of the isoquant to the upper starting point labour input is reduced by 11.1%, while the percentage reduction in energy consumption by moving from the upper starting point to the lower end point of the isoquant is 9.6%.

6

Empirical Analyses: An Overview

6.1 Introduction

In Chapters 1–5 we presented a theoretical basis and estimation methods for the empirical analysis of industrial structure and structural change. In this chapter there is a discussion of how we applied this approach to the empirical analyses of several industries. Our purpose is to examine various methods of extracting information from a given set of data and to illustrate various ways of looking at the data in order to grasp the nature of the structure.

As pointed out in Chapter 1, an analysis of industrial structure requires a dynamic theory of production. Various models can be formulated depending on the degree of inertia allowed in the capital structure. One of the most important models generating stability and inertia in the capital structure is the putty-clay model, which was discussed in Chapter 2. Our production function framework is based on Johansen's distinction between the *ex ante* and *ex post* functions at the micro level and the distinction between the short-run and long-run production functions at the industry level. These concepts were introduced in Section 1.4.

In order to develop a comprehensive long-run analysis of technical progress and structural change information about both the short-run function and the ex ante micro functions is required. As discussed in Chapter 4, the ex ante function can be derived from engineering knowledge or estimated as a frontier production function. The former case requires considerably more information about the technical relationships.

The estimation of a frontier production function depends crucially on capital data. In addition, the identification of vintages of different technologies is required to give empirical content to the putty-clay model. In view of the difficulties in obtaining such data, it should be noted that the data requirements for the short-run function are limited to the current inputs by the very nature of this production function concept.

It has not been possible to integrate all the different production function concepts for any one of the industry analyses. Instead, we used the frontier production function concept in the analysis of one of the industries and the short-run industry production function in the analyses of the others. These functions may be seen as useful tools for the analysis of industrial structure and structural change within a putty-clay framework, even though a total integration has not been obtained. We studied structural change by examining the development of the frontier and the short-run function over time. Although the putty-clay model forms our basic foundation, we do not formally test whether this model is tenable,[1] rather we justify the assumption by use of engineering knowledge.

The empirical analyses presented here are based on different sets of data for Swedish and Norwegian industries or industrial activities. The data comprises milk processing activities, cement kilns, pulp plants, blast furnaces and aluminium plants.

The data differ in nature. In two cases, cement kilns and blast furnaces, the data refer to the central piece of capital equipment within a plant. In other cases we have obtained data for the entire plant as the basic micro unit, or for a specific activity such as general milk processing.

For all industries except milk processing we have obtained reliable data for both labour and energy inputs. On the other hand we have not succeeded in obtaining reliable data for capital equipment. Thus, these data sets are suitable for short-run industry production function analysis. For general milk processing we have data on capital and labour. Two other important inputs are raw milk and energy. The former is strictly proportional to output while the latter is closely related to output, its basic use being for heating and cooling according to prespecified standards common to all dairies. Since the dairies utilised very different energy sources (e.g., saw dust, wood chips, oil, water, etc.) and only registered energy costs, it has not been possible to construct physical energy data. The analysis, therefore, concentrates on a frontier production function which reveals technical change with respect to capital and labour.

[1] See Fuss [1977, 1978].

6.2 Description of structure

Introduction

The concept of structure was defined in Section 1.3. The elements of structure to which we now pay particular attention are the distribution of input coefficients (input per unit of output) and the capacity and output of the micro units of the industry.

Partial input-coefficient distribution

Our starting point is how the individual units utilise their inputs. Descriptions of structure should show the distribution of input utilisation over units. One way of organising the data is to look at the distribution of input coefficients for one factor at a time. Measuring the input coefficient along the ordinate axis and absolute or relative level of production along the abscissa axis, we may enter the units in order of increasing value of the input coefficients, as for example in Figure 6.1.

Each bar in the histogram represents a unit. Such a figure of an input-coefficient distribution may be termed a Salter-diagram.[2] Measuring average costs along the ordinate axis turns it into a Heckscher-diagram as defined in Chapter 2. This way of organising the data gives us directly the following information:

(i) the range of variation in the input coefficients

(ii) the form of the input coefficient distribution

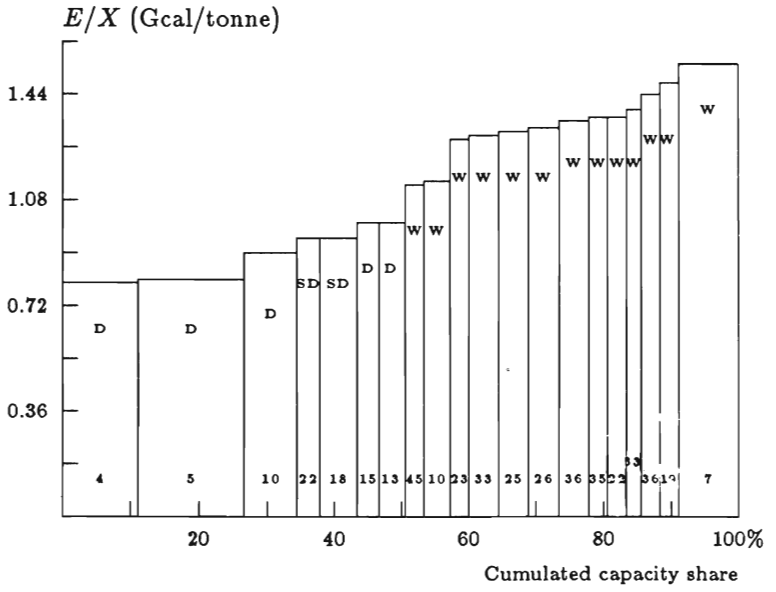
(iii) the relationship between the input coefficient distribution and the size distribution of the units.

If we have additional information about the age of each unit, this information may also easily be entered into the figure, giving us a picture of the relationship between the input-coefficient distribution and the age distribution, as can be seen in Figure 6.1.

The development of the input-coefficient distribution through time can be studied by plotting the surface of the histograms for different years as in Figure 6.2. Such a diagram reveals:

(i) changes in the form of the distribution

[2] See Salter [1960].



D, SD and W denote dry, semi-dry and wet kilns, respectively.
The figures at the base of the histogram denote the ages.

Figure 6.1: Energy input-coefficient (E/X) distribution for the Swedish cement industry in 1974.

(ii) changes in the range of the input coefficients, i.e., the difference between the greatest and the smallest value

(iii) changes in the position of small and large units.

Partial input-coefficient distributions may also be represented, as in Sato [1975], by changing a Salter-diagram to measure the sorted input coefficients along the abscissa axis and capacity shares (not cumulated) along the ordinate axis.[3] Such a representation may be better suited to aggregated data.

If one is interested in approximating a *continuous* capacity distribution by means of partial input-coefficient distributions, one should adjust representations such as those mentioned above to take account of the fact that observed capacity within an interval on the input-coefficient axis is

[3] See Sato [1975], p. 164.

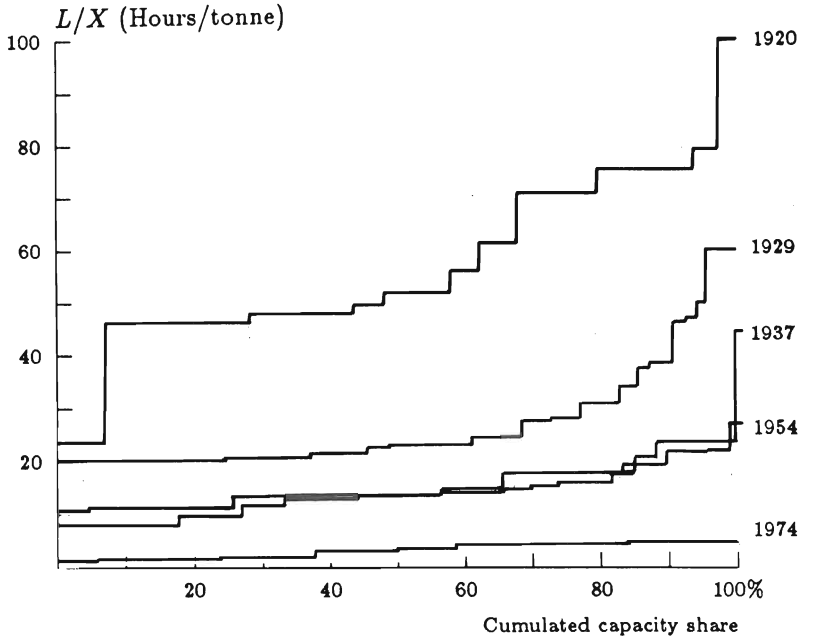


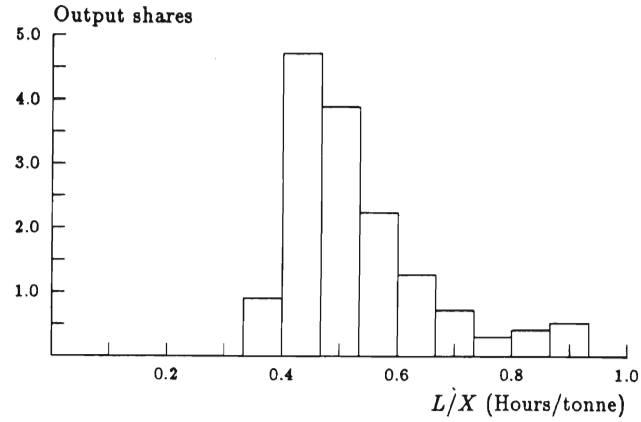
Figure 6.2: The development of the labour-input coefficient (L/X) distribution between 1920 and 1974 for the Swedish sulphate pulp industry.

equal to the integral under the continuous capacity distribution over the same interval.[4] When representing such a partial distribution the capacity observed within an interval should be distributed over the interval length. This has been carried out in Figures 6.3 and 6.4. In addition the areas have been normalised to 1 by dividing through by total capacity. Entering such partial distributions for different years indicates type and range of change, as illustrated in Figure 6.4.

Another way of looking at the development of the input coefficients so as to gain a more summary picture of the process of structural change is by comparing the average values of the input coefficients and the best practice values of the input coefficients.[5]

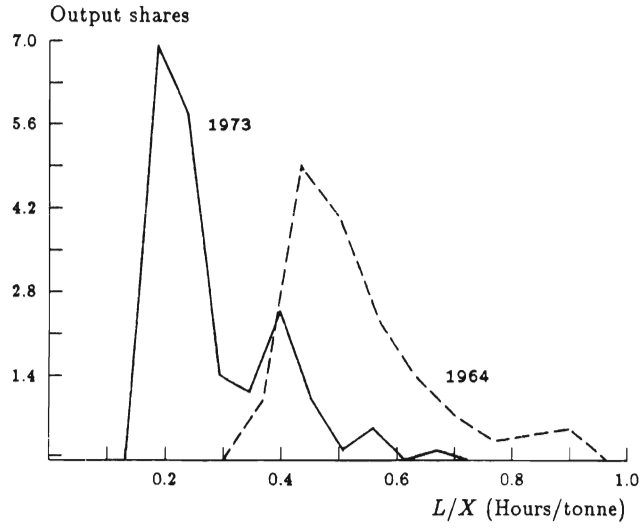
[4] See Muysken [1979, 1983, 1985].

[5] See Maywald [1957]. For an application, see Figures 7.1 and 7.2 in Chapter 7.



Area of the histogram represents output shares.

Figure 6.3: Output shares within intervals of labour-input coefficient (L/X). Swedish dairies, 1964.

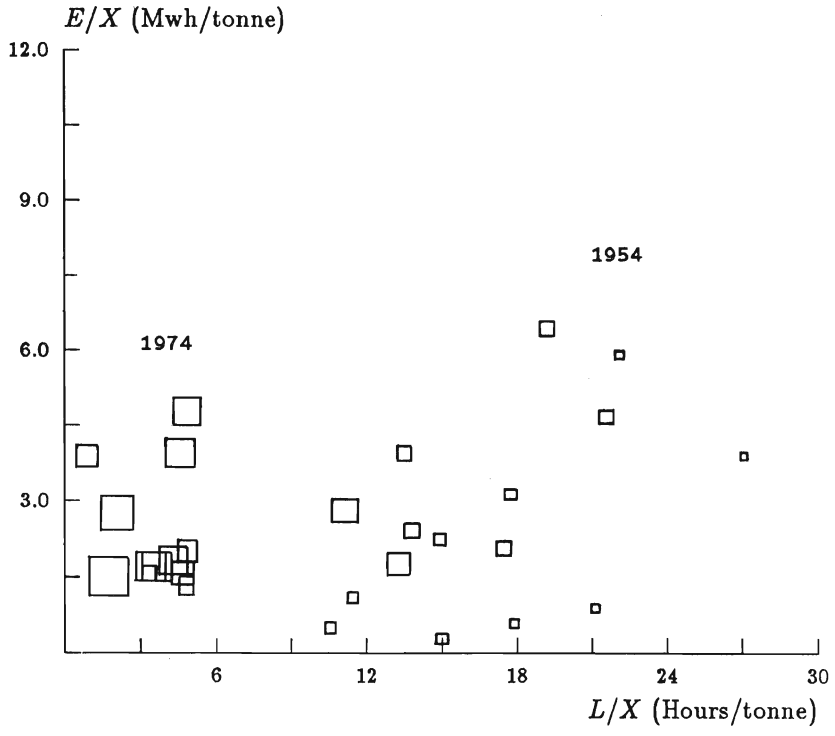


The dashed curve represents 1964 and the solid curve 1973.
The area of the histogram is erected around the aggregated observations represented by the kinks in the curves.

Figure 6.4: Output shares within intervals of labour-input coefficient (L/X). Swedish dairies, 1964 and 1973.

Capacity distributions in the input-coefficient space

The Heckscher-Salter-diagrams give a partial description of structure. In order to get a more comprehensive description of the inputs used a simultaneous presentation is needed. Graphically such a description can only be given for three inputs at a time. We will limit ourselves to two-dimensional diagrams. The input coefficients are measured along the axes and the level of production or capacity for each unit may, for instance, be illustrated by representing each observation as an area in the form of a square, circle, etc., proportional to the level of production or capacity or the unit's share of total sector output or capacity. An example is shown in Figure 6.5.



The sizes of the squares are proportional to the capacity.

Figure 6.5: Capacity distribution diagram in the input-coefficient space: Swedish sulphate pulp industry in 1954 and 1974.

This kind of diagram is called a capacity distribution diagram.[6] The diagram brings out how the sector's production capacity is distributed with respect to the two input coefficients.[7] A capacity distribution diagram combines information from two Salter-diagrams. A capacity diagram can be read in the same way as a Salter-diagram. The range of variation is shown for both types of input coefficients. With respect to the shape of the capacity distribution it is of interest to see, for instance, if the capacity is in a southwest-northeast direction, or in a southeast-northwest direction. It is especially illuminating to look at the changes in structure between two different points in time, as can for example be seen in Figure 6.5.

If structural change is revealed graphically to be the result of technical progress, i.e., the need for both input coefficients has decreased simultaneously, this may be due to either neutral technical change or an increased exploitation of economies of scale. If a structural change is characterised by a graphic transformation of the structure in a northwest-southeast direction, two possible explanations are:

(i) the development of the ex ante or choice of technique production function

(ii) the development of relative factor prices and its influence on scrapping and the choice of technology in new equipment.

It may be the case that the number of units in the industry is so large that aggregation of units have to be done in some meaningful way in order to utilise the diagrams developed above. Dependency on primary data sources restricted by secrecy codes might compel a certain level of aggregation too.

As an empirical illustration of the use of capacity distribution data Johansen examined 377 Norwegian tankers in 1967.[8] Dividing the observed range of energy and labour-input coefficients into 20 intervals, the resulting capacity distribution based on an aggregation within each cell is shown in Figure 6.6.[9] In addition the capacity region of the short-run industry function is shown[10] starting at the most efficient unit and ending, obviously, at the centre of gravity of the capacity distribution, thus being considerably more concentrated than the latter. The branching out of the capacity

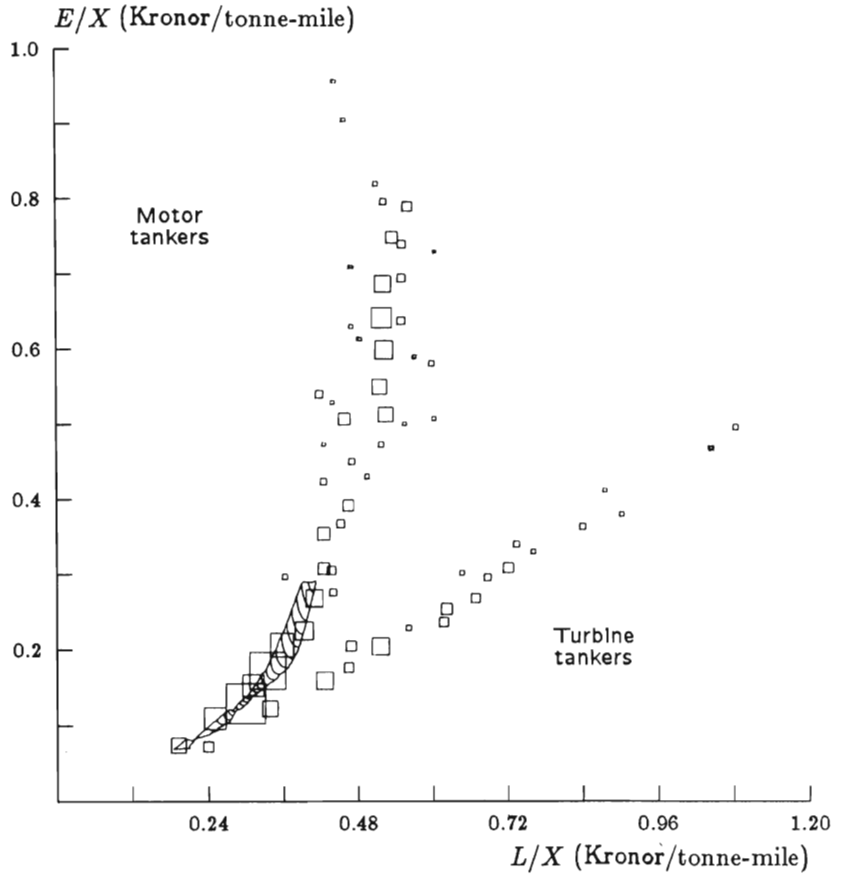
[6] See Johansen [1972], p. 247.

[7] Sato [1975] calls this diagram the efficiency distribution.

[8] See Johansen [1972], Ch. 9.

[9] This consistent method gives a slightly different picture than shown in Johansen [1972], Figure 9.1, p. 247.

[10] With the same number of isoquants as in Johansen [1972], Figure 9.2, p. 256.



The sizes of the squares are proportional to the capacity.

Figure 6.6: The capacity distribution and capacity region for Norwegian tankers, 1967. Labour-input coefficients (L/X) and energy-input coefficients (E/X).

distribution into an upper branch consisting of motor tankers and a lower of turbine tankers is reflected in a wider capacity region at the upper end. Since motor tankers constitute the dominating part of total capacity, the capacity region follows the shape of the motor tanker branch.

6.3 The main empirical results

Frontier production functions

Chapter 4 is concerned with the estimation of a frontier function based on observed performances. This frontier could be referred to as the Best-Practice function.[11]

As pointed out in Section 4.2, the key question when defining the frontier function concept is whether to allow actual observations to be above the frontier or not. The frontier is called deterministic if all the observations are required to lie on or below the frontier and stochastic if observations are allowed to be above the frontier due to random events.

With respect to estimation procedures for deterministic frontiers the main issue is whether the efficiency differences between the units are assumed to be generated by an explicit efficiency distribution or not. In the latter case the frontier must be computed in a more or less arbitrary way, while in the former case it is, in principle, possible to derive maximum likelihood (ML) estimates.[12] An empirical analysis of frontier functions is presented in Chapter 7.

The analysis is based on a panel set of cross-section time-series data for 10 years, 1964–73, of 28 individual dairy plants producing a homogeneous product, dairy milk, with inputs capital and labour. Estimation of production functions on the basis of time-series data is usually carried out at a very high level of aggregation. Cross-section data on individual plants producing a homogeneous output are rather scarce except in the field of agriculture and electricity generation.[13] The analysis in Ringstad [1971], however, is based on pooled time-series cross-section data, but the level of aggregation is rather high, since the base unit of the industry construction is the two-digit group. Earlier studies have almost exclusively been limited to estimating Hicks-neutral technical progress in production functions fitted as an average of the sample. Exceptions here are Ringstad [1974], Sato [1970] and Greene [1983], who studied non-neutral technical progress.

In the present study technical progress is analysed by introducing trends in all the parameters of the frontier production function. In particular, trends are introduced in both of the scale function parameters, thus

---

[11] See Salter [1960].

[12] The former approach is followed in Afriat [1972] and Schmidt [1976].

[13] See, e.g., Christensen and Greene [1976], Dhrymes and Kurz [1964], Komiya [1962] and Nerlove [1963].

making it possible to study whether the optimal scale changes over time. To further elucidate the process of technical advance, Salter's measures of technical advance have been generalised, in a way inspired by Farrell.[14]

The frontier function is first estimated without assuming an explicit efficiency distribution. Without an explicit efficiency distribution it seems natural that the objective be to have the observations as close to the frontier as possible in some sense. In Aigner and Chu [1968] both the sums of simple and squared deviations from the frontier were used. In order to keep the estimation problem as simple as possible, we have chosen to minimise the simple sum of deviations from the frontier with respect to input utilisation constraints. With this specification the estimation problem is reduced to the most simple problem of solving a standard linear programming problem with the homothetic functional specification chosen.

For the main empirical results, when allowing variable returns to scale, the driving force behind technical progress turned out to be a fairly rapid shift in the returns-to-scale function. The upward shift of the production frontier tended to be non-neutral, increasing the kernel elasticity of labour and decreasing the kernel elasticity of capital somewhat.

The splitting up of the generalised Salter measure shows that it is the movement of the efficiency frontier along a ray towards the origin that results in the significant reductions in the average costs at the optimal scale of 9–13 percent per year. Optimal adjustment to the capital saving bias results in quite insignificant cost reductions.

The sensitivity analysis revealed that the production function parameters were influenced by the *a priori* discarding of chosen units, some of which turned out to be on the frontier of the complete sample. However, the form and shift of the elasticity-of-scale function were fairly stable, leading to quite small variations in the cost reduction measures.

Farrell's measures of productive efficiency were elaborated and generalised to non-homogeneous production functions in Section 3.4. These new measures of efficiency have been applied to the Swedish milk processing industry. The development of the industrial structure is studied by the change in the efficiency distributions for the individual plants through time. The aggregate performance of the sector is studied by examining the development of the different measures of structural efficiency.

The most remarkable result is the rather large distance between best-practice and average performance measured by different measures of structural efficiency. Moreover, this distance shows an increasing trend during

[14] See Salter [1960], Chapter 3.

the period. These results are explained by rapid technical progress in combination with an underlying putty-clay technological structure and a slow growth of investment.

The distribution of the individual measures of technical efficiency and scale efficiency reveals a large variation in efficiency between the units for all years. Some of these differences in efficiency can be explained by the modernity of equipment and others by differences in management capability. The basic methodological problem with the deterministic computational approach is that the tools of statistical inference do not apply.

One development in the estimating of frontier production functions has been the introduction of a composed error structure in the production function to allow simultaneously for systematic efficiency differences between production units and random differences. The purpose of Section 7.4 is to compare the results obtained with this specification and the previously developed techniques for estimating deterministic frontiers. The estimations are carried out on cross-section data. Results for the pooled data-set are also given and a number of structural efficiency measures are computed.

Comparing the results of the deterministic and the stochastic approaches to estimating the frontier production function, we find that the parameter values for each year differ considerably. In the deterministic case we also obtain that they behave in an unsystematic way from year to year. This is also the case for the technical optimal scale output levels. The parameters estimated from the composed error model (ML-CE) are more stable than the others. At the same time the difference between the ML-CE results and the results of the corresponding average production function model is very small. The ML-CE production frontier tends to be a neutral shift of the average production function. The sensitivity analysis reveals that removing one frontier unit in the linear programming (LP) case has considerable effect on the stability of the results, without moving the LP frontier in the direction of the average production function. The LP frontier is even more stable from year to year than the average-like ML-CE frontier.

By the nature of the LP and ML estimation procedures one would expect more or less strong differences in the estimated parameters between the different years when, as is the case here, both the input coefficients and output levels and the set of on-the-frontier observations gradually change from year to year. When assessing frontier estimation one must keep in mind that the very purpose of frontier function estimation is to count the most efficient units disproportionately if the data for these units are trust-

worthy. In our case, the data have been checked carefully, and there are no extreme outliers which have abnormally low values of input requirements.

According to the ML-CE approach one would expect the outliers, i.e., the units close to the LP and ML frontiers, to vary randomly from year to year. However, when looking at the most efficient units in both the LP and ML cases there turns out to be a high degree of stability between consecutive years; especially high for the smallest and the largest units in the LP case as was pointed out above. Units with intermediate output levels are usually close to the frontier about 2–4 years consecutively. Thus, the overall impression is a high stability from year to year but a gradual change during the whole period.

The estimates of structural efficiency corresponding to the LP and ML estimations are significantly lower than the ML-CE values. For an industry with long-lasting equipment and a rather rapid technical progress one would expect large differences in input requirements between the different units, i.e., between best-practice and average performance of the industry.

Short-run industry functions

As outlined in Chapter 5, the short-run industry production function is established by maximising output for given levels of current inputs. The short-run function gives a unique relationship between the actual technology of the individual units and the short-run industry function. Both structural change and technical progress are revealed by utilising the short-run industry function. The realised production at any point of time must be compatible with the short-run industry production function. A study of the dynamics of the production of a sector requires a study of how the short-run production function changes over time.[15]

The purpose of Chapters 8 to 11 is to provide a deeper empirical insight into the structural change of an industry. The main contribution is a long-run analysis of technical progress and structural change by means of the short-run industry production function.

The following three aspects of technical change are studied empirically:

(i) factor bias, i.e., shift of the substitution region

(ii) productivity change, i.e., shift of the isoquants towards the origin

(iii) changes in the shape of the isoquants, i.e., change in substitution properties.

[15] See Johansen [1972], p. 26.

To further elucidate the process of technical advance we have also generalised, in a way inspired by Farrell, Salter's measure of technical advance for this type of production function.

In comparison with the high-brow econometrics of empirical production theory, our approach may seem less sophisticated.[16] On the other hand it yields a deeper insight into the nature of the development of an industry

In Chapter 8 the short-run function approach is applied to an empirical analysis of technical progress and structural change in the Swedish cement industry during a twenty-five year period, 1955–79. The analysis is based on micro data for individual kilns.

The empirical results show that the process of structural change of the Swedish cement industry has been characterised by a substitution process from labour towards energy in combination with a rather rapid cost reducing technical progress. This development is due to long-run ex ante substitution possibilities between capital and labour/energy and increasing returns to scale when introducing new techniques, and disembodied improvements especially as regards labour saving.

The Swedish pulp industry is analysed in Chapter 9 spanning a considerably longer period than for the cement industry. The period 1920–74 is covered by five cross-section data sets on the three processes: sulphate, sulphite and mechanical pulp. There is an overall labour saving bias together with a cost-reducing technical change. But this development was abruptly reversed during the years of the second world war when both *labour-using bias* and *cost-increasing technical change* occurred. (A more detailed analysis reveals the same reversed development during the extreme Korea boom years 1951–52.)

An even longer time span is investigated for Swedish pig iron production in Chapter 10, covering the years 1850, 1880, 1913, 1935, 1950 and 1974. Technical change has been different in scope and nature between the various periods, starting with an overall technical progress that was the result of the average catching up with best-practice technology. In the next two periods technical progress took place due to the introduction of new technologies and increasing unit size.

While the pulp industry lost its international markets during the war, the steel industry carried on at an uninterrupted production level, showing about the same degree of labour saving bias and technical progress as during the prewar period. In the postwar period there has been a particularly

[16] Cf., Johansen [1972], p. 1.

strong labour saving bias.

The Norwegian primary aluminium industry is studied in Chapter 11 for about the same time span as the cement industry in Chapter 8. The production unit is an establishment, and current inputs electricity and labour. With respect to the empirical results in Chapter 11, there has been a marked shift of the substitution region towards the electricity axis. Direct substitution between electricity and labour is possible only to a very limited extent when capital is a variable factor. Thus we interpret the result as clear evidence of labour-saving technical change over the period of observation, probably induced by the rise in the relative price of labour compared to electricity and the technical possibilities for cost reductions. *Hicks-neutral technical change* is thus not supported by data for the aluminium industry.

The industry production function for aluminium is characterised by narrow substitution regions for all years, reflecting a high degree of technical uniformity between Norwegian aluminium smelters. The straight and narrow regions of substitution indicate further that the short-run production function of the aluminium industry can be adequately represented by a simple Leontief function.

7

The Swedish Dairy Industry

7.1 Introduction

In this chapter we present an empirical investigation of a part of the Swedish dairy industry, a study for which there was available very reliable micro data. In particular, the data set that we used was suitable for the following three analyses:

(i) a comparison of methods for the estimation of frontier production functions

(ii) an analysis of technical progress on the basis of frontier production functions

(iii) the measurement of productive efficiency.

The processing of milk in a dairy can be divided into different stages. Each stage can be referred to as a production process. The data used in this study refer to one such production process, namely, general milk processing. This process includes the receiving of milk from cans or tanks, storage, pasteurisation and separation. All the milk passes through this process before it goes further on to other processes which ultimately lead to market milk, butter, cheese, milk powder, etc. Hence, this stage defines the capacity of the plant. General milk processing is often treated as a separate unit in cost accounting. Moreover, the Swedish Dairy Federation collects yearly data for all the different processes mentioned above. As the data are separated for the different processes it is possible to analyse each step in the dairy operation individually.

A strong reason for our choice of the general milk processing stage is that it enables one to measure output in physical or technical units (tonnes) avoiding value added or gross value of output. This means that our estimated production function is more of a technical production function in the original sense.

7.2 Data

In the empirical part of this study we have utilised primary data for general milk processing from 28 individual dairy plants for the period 1964–73. We received all our data from the Swedish Dairy Federation, SMR, a central service organisation for the dairies in Sweden.

With respect to the reliability of the data it should be mentioned that the very purpose of the data collection by SMR is to measure efficiency. While the labour figures are reported by the dairies themselves, the capital figures are very carefully calculated by SMR. The equipment and buildings of the dairies are controlled at regular intervals.

In the study milk is regarded as a homogeneous product, a realistic assumption. Output is measured in tonnes of milk delivered to the plant each year. The amount of milk received is equal to the amount produced. There is no measurable waste of milk at this stage. According to SMR any difference is due to measurement errors. (Differences are of the magnitude of kilos.) Moreover, there is no potential substitutability between raw milk and other inputs.

The labour input variable is defined as the hours worked by production workers including the technical staff, which usually consists of one engineer.

Capital data of buildings and machinery are of user-cost type, including depreciation based on current replacement cost, cost of maintenance and rate of interest. The different items of capital are divided into five different subgroups depending on their durability, which varies between 6 and 25 years. Hence the capital measure is an aggregated sum of capital costs from these subgroups.

Capital costs, divided into buildings and machinery, are calculated on the basis of these subgroups. The capital measure has been centrally calculated by SMR, according to the same principles for all plants and after regular capital inventory and revaluations by engineers from SMR. Afterwards we have aggregated buildings and machinery into one capital measure. In view of the composite commodity aggregation theorem it is

interesting to note the fact that the relative prices of buildings and machinery have developed almost proportionally during the 10-year period. The price index has moved from 100 in 1964 to 158 in 1973 for buildings and to 161 for machinery. An alternative would have been to retain the disaggregation of buildings and machinery, however, in the case of a C-D kernel function this would imply a unitary elasticity of substitution. This seems to be a less realistic assumption, however. Note that this capital measure is proportional to the replacement value of capital, which may serve as a measure of the volume of capital.[1]

Since the data are not adjusted for capacity utilisation, we have investigated a measure based on the monthly maximum amount of milk received compared with the yearly average. This ratio is fairly stable over time, and the differences between the plants are quite small. Since we are not sure that this capacity measure reflects the real capacity concept, and as we know it is almost proportional to the current output figures, we have consequently not corrected current output.

We will employ the following notations:

x = quantity produced milk in tonnes
L = working hours by production workers
K = user cost of capital in Swedish kronor (1964 prices)
N = number of units
T = number of years

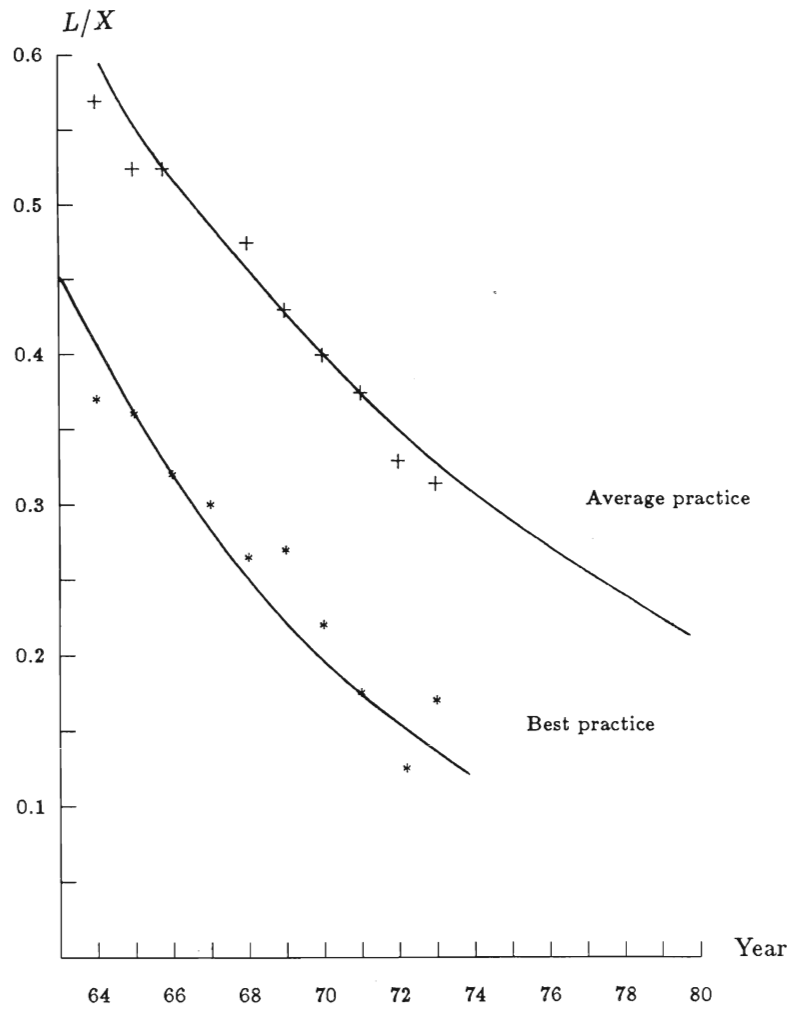
7.3 Structural description

In Figures 7.1 and 7.2 below the development of the observed input coefficients are plotted together with log linear regression curves of the development over time. The crosses denote the average values of the input coefficients and the stars the best practice values for each year during the period.[2]

The difference between the average and best practice coefficients has been about the same over the period, while the number of years during

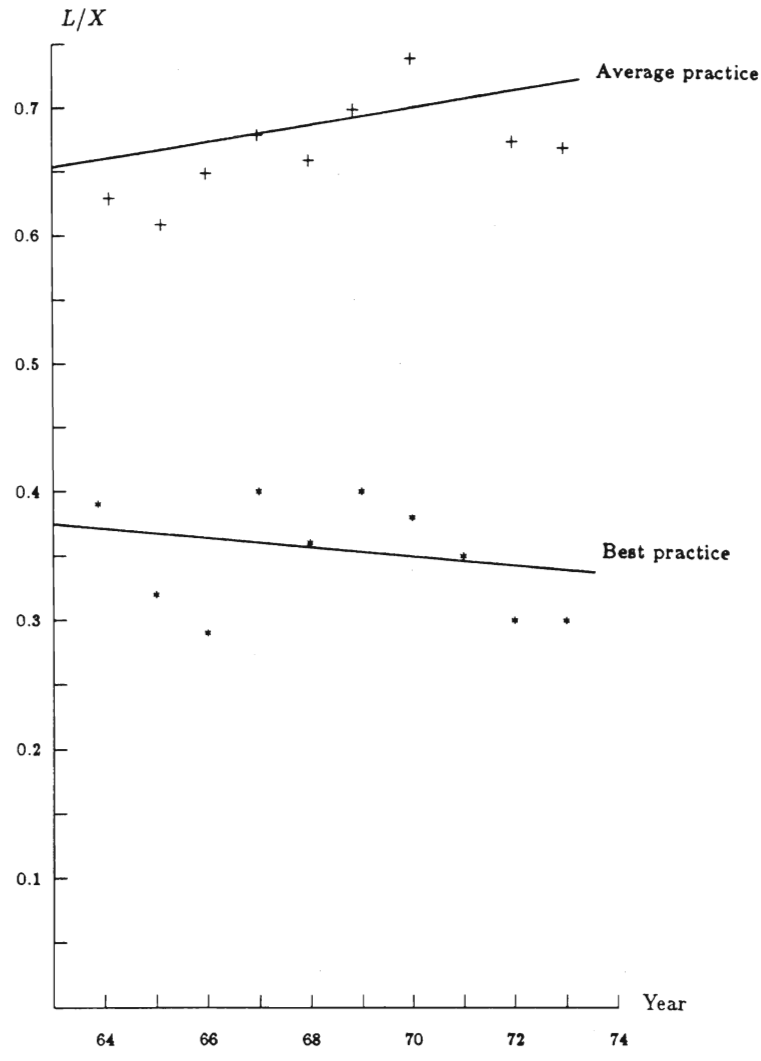
[1] See Johansen and Sørsveen [1967].

[2] See Maywald [1957].



Observations for working hours per tonne milk: the (arithmetic) average (crosses) and best-practice plants (stars) for each year 1964—1973, together with the plotting of log linear regressions for the development over time.

Figure 7.1: Observed best practice and average input coefficients for labour.



Observations for capital (1964–kronor) per tonne milk: the (arithmetic) average (crosses) and best-practice plants (stars) for each year 1964—1973, together with the plotting of log linear regressions for the development over time.

Figure 7.2: Observed best-practice and average input coefficients for capital.

which the average has lagged behind the best practice coefficient has increased from approximately five years in 1964 to 11 years in 1973. For capital, both average and "best practice" coefficients have been almost constant. Best practice should not be interpreted normatively in such a partial setting, since factor substitution must be behind the development shown in the figures.

The Salter diagrams for the years 1964, 1968 and 1973 are shown in Figures 7.3 and 7.4. The value of the input coefficient in question is denoted on the ordinate axis, and each individual unit is represented by a step, the width of which is proportional to the output of the unit in question.

The distributions are fairly flat up to about 60 per cent of the total output, and after that they increase more markedly. As regards the location of small and large plants there are relatively more large units with low input coefficients and more small units with large input coefficients.

The figure demonstrates the parallel downward shift in the labour input-coefficient distribution. We can see that the relative position of the largest plant has deteriorated, especially in the last period from 1968 to 1973. In Chapter 6, Figure 6.4 gives a summary picture of the development from 1964 to 1973. Both the change in the range and the shape of the distribution are revealed. The more even cumulated distributions for 1973 in Figure 7.3 show up as a higher concentration of capacity close to the best-practice level in Figure 6.4. The second peak in 1973 in Figure 6.4 is due to the relatively high labour-input coefficient of the largest unit.

Figure 7.2 illustrates that both average and best-practice input coefficients of capital have been stable over time. Figure 7.4 clearly shows that the whole distribution has been remarkably stable. The partial input-coefficient diagrams are put together in the capacity distribution diagrams and are presented in Figures 7.5 and 7.6.

With respect to structural change from 1964 to 1973 a fairly wide distribution in the northwest-southeast direction has been changed to a distribution located in the northeast-southwest direction. At the same time the distribution has moved towards the origin and the capital axis. The large plants have on the whole smaller input coefficients. In 1964 the largest plant was fairly efficient, while in 1973 it was located in the middle of the distribution, and has a higher labour input coefficient than the other large units.

Since the output scale of the squares is the same for both years, Figures 7.5 and 7.6 also give a picture of the development of the size distribution. The largest plants have grown larger and the smallest plants smaller. The average output has increased by about 30 percent.

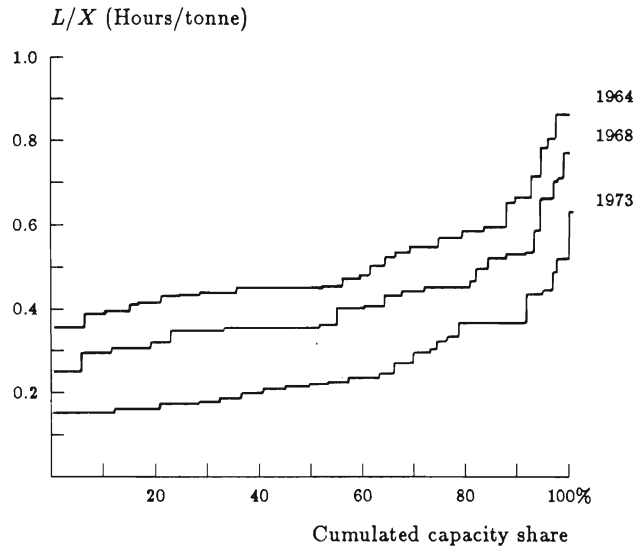


Figure 7.3: Salter diagram: Labour-input coefficients, 1964–68–73.

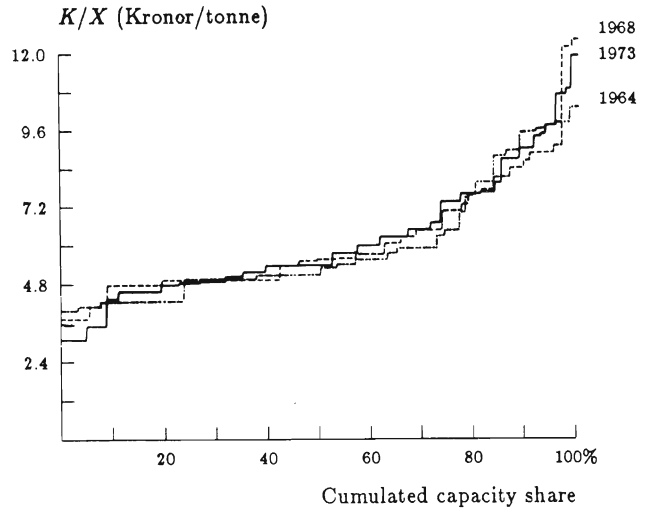


Figure 7.4: Salter diagram: Capital-input coefficients, 1964–68–73.

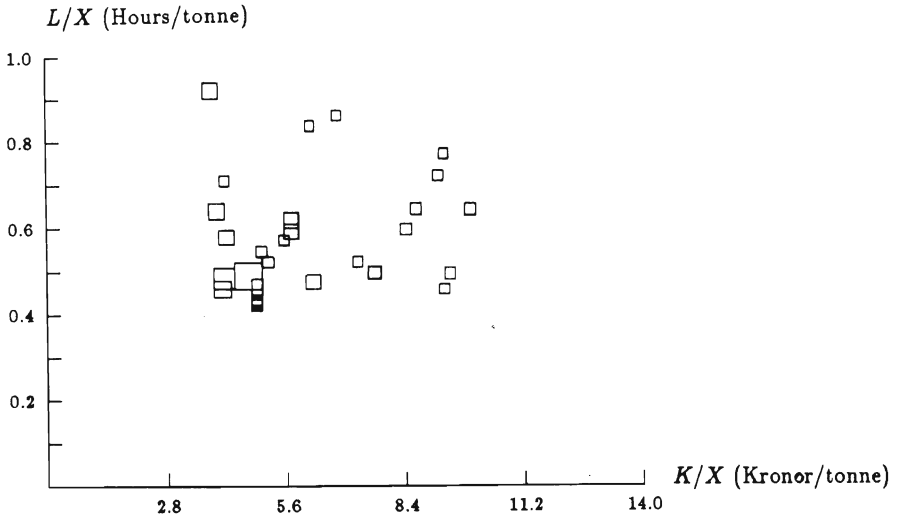


Figure 7.5: The capacity distribution of 28 Swedish dairies in 1964.

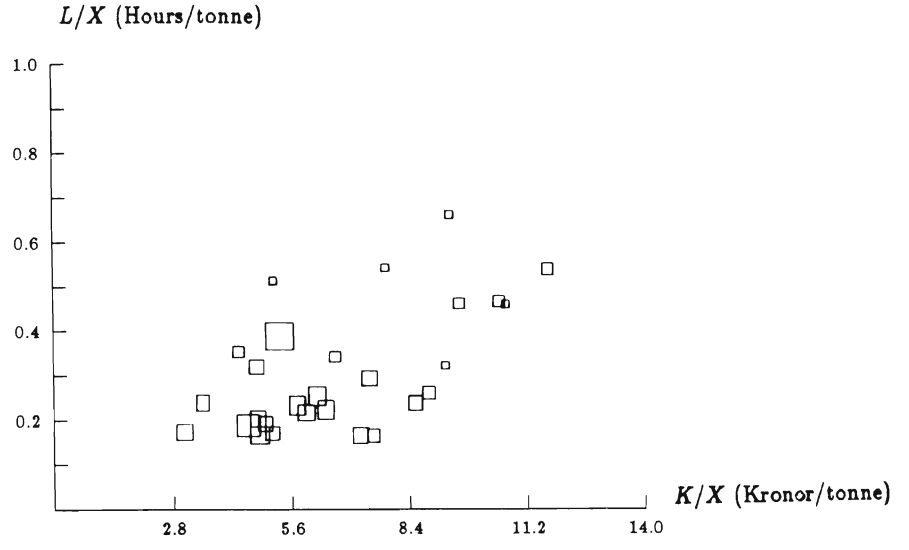


Figure 7.6: The capacity distribution of 28 Swedish dairies in 1973.

7.4 Estimation of deterministic and stochastic frontier production functions

Introduction

In this section we examine three of the different ways of estimating frontier production functions that were presented in Chapter 4:

(i) computing the frontier by solving an LP-problem with on-or-below-frontier constraints

(ii) introducing an explicit efficiency distribution and deriving ML-estimates by solving a non-linear programming problem

(iii) using a composed error (CE) term, the first part of which represents the differences between the units due to inefficiency, and the second part random disturbances generated by measurement and specification errors and random events in the real sense.

The different methods and the applied functional forms are treated in Chapter 4. The functional form used here is that presented in Section 4.2:

$$x^{\alpha}e^{\beta x} = A\prod_i v_i^{a_i}\,, \qquad \sum_i a_i = 1 \tag{7.1}$$

Cross-section results

The main results are set out in Tables 7.1 and 7.2. (The columns denoted by LP* show the results of a sensitivity test and will be commented upon later on.) Comparing the parameters computed by solving the linear programming problem (LP) with the values obtained from the maximum likelihood estimates with an exponential efficiency distribution (ML) and a composed error term (ML-CE), we find a rather systematic pattern of differences.

The kernel elasticity of labour is generally higher for the LP-computations than for the ML-CE-estimations, and vice-versa for the capital kernel elasticity. The ML-results tend to be close to the LP-results.

The last row of Table 7.1 shows the standard deviations around the mean of the observed parameter values over time. In general, the ML-CE-results show a lesser variability than the LP and ML results.

In order to further illustrate the differences, the graphs of the production functions along the ray of the average factor proportion are illustrated

Table 7.1: Comparison of the results for the frontier production function $x^{\alpha}e^{\beta x} = AL^{a_L}K^{a_K}$. x is output, L is labour input and K capital input.

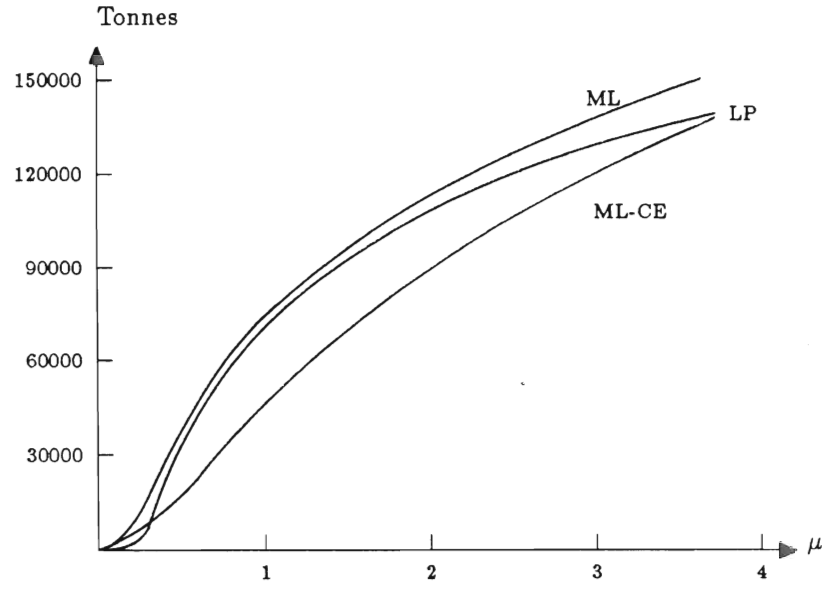
Year	Constant, $\ln A$				Labour kernel el., a_L				Capital kernel el., a_K			
	LP	LP*	ML	ML-CE	LP	LP*	ML	ML-CE	LP	LP*	ML	ML-CE
1964	−4.64	−5.22	−4.34	−2.10	.68	.58	.70	.64	.32	.42	.30	.36
1965	−9.07	−4.40	−3.35	−.86	.76	.65	.45	.65	.24	.35	.55	.35
1966	−5.62	−6.82	−1.68	−2.01	1.	.68	.96	.58	0.	.32	.04	.42
1967	−7.56	−7.67	−7.55	−3.50	.55	.69	.54	.55	.45	.31	.46	.45
1968	−4.84	−6.61	−7.54	−3.65	.64	.63	.61	.60	.36	.37	.39	.40
1969	−4.24	−5.56	−3.76	−4.10	.65	.65	.78	.58	.35	.35	.22	.42
1970	−6.93	−3.61	−.52	−2.68	1.	.58	.9999	.43	0.	.42	.0001	.57
1971	−6.56	−5.74	7.43	−4.60	1.	.75	.9999	.49	0.	.25	.0001	.51
1972	−4.27	−6.26	−3.72	−5.32	.72	.72	.9999	.60	.28	.28	.0001	.36
1973	−8.62	−5.49	−5.01	−4.82	.81	.53	.87	.51	.19	.47	.13	.49
Mean	−6.24	−5.74	−3.00	−3.30	.78	.65	.79	.56	.22	.35	.21	.43
Std dev	1.79	1.18	4.28	1.43	.17	.07	.21	.07	.17	.07	.21	.07

LP* is the sensitivity test. The unit with the highest shadow price (on the constraint (4.7)) each year is removed.

Table 7.1 continued:

Year	Scale function parameter, α				Scale function parameter, $\beta \cdot 10^5$			
	LP	LP*	ML	ML-CE	LP	LP*	ML	ML-CE
1964	.52	.47	.54	.81	.88	1.13	.84	.16
1965	0.	.55	.68	.93	2.4	.80	.91	−.08
1966	.31	.28	.72	.83	1.3	1.25	.71	.19
1967	.24	.19	.24	.68	1.26	1.35	1.26	.34
1968	.52	.32	.23	.65	0.	1.01	1.01	.35
1969	.58	.43	.59	.61	0.	.82	.02	.43
1970	.14	.63	.77	.79	1.7	.67	.93	.23
1971	.18	.34	1.52	.51	1.0	1.26	.26	.72
1972	.52	.29	.44	.45	0.	1.37	1.33	.90
1973	.003	.43	.34	.53	1.9	.93	1.33	.79
Mean	.30	.39	.61	.68	1.04	1.06	.86	.40
Std dev	.22	.13	.37	.16	.84	.25	.44	.31

LP* is the sensitivity test. The unit with the highest shadow price (on the constraint (4.7)) each year is removed.



The production function cut with a vertical plane through the origin along the ray $(\mu L^0, \mu K^0)$ of average factor proportion
$x^{\alpha} e^{\beta x} = A(\mu L^0)^{a_L} (\mu K^0)^{a_K}$

Figure 7.7: The frontier production functions for 1973.

in Figure 7.7 for the year 1973, which is regarded as typical for the series of cross-section results.

The ML-frontier represents an upward shift of the ML-CE-frontier, but is more curved than the ML-CE frontier. The LP-frontier is even still more curved, intersecting the ML-CE-frontier both for small and large values of output.

For medium-range output levels (between 40,000 and 75,000 tonnes) and for given amounts of inputs, both the LP and ML-frontiers yield considerably higher predictions of output levels than the ML-CE-frontier, but for higher output levels the LP-frontier approaches the ML-CE-frontier. It turns out that almost all the units on or close to the frontier in the ML-case are also on the frontier or very near the frontier in the LP-case. On the other hand, the frontier units in the ML-case are seldom dispersed over the whole range of observed output levels but are usually concentrated within the medium range and one of the tails. Usually, in the LP-case, one of the smallest and one of the largest units are on the frontier together with one

Table 7.2: Comparison of optimal scale in tonnes of milk, $\hat{x} = (1-\alpha)/\beta$, of the frontier function $x^{\alpha}e^{\beta x} = AL^{a_L}K^{a_K}$.

Year	Average output level	Optimal scale output levels			
		LP	LP*	ML	ML–CE
1964	28732	54545	46866	54722	121675
1965	28970	41667	55841	35321	**
1966	31145	53076	57333	39725	87662
1967	30912	60839	60098	60990	92396
1968	34501	1.92[1]	66868	76323	98856
1969	35008	1.72[1]	69987	1929571	90978
1970	35171	50588	55490	24622	93116
1971	36154	82000	52242	*	68034
1972	36708	1.91[1]	51803	42570	61334
1973	38807	52473	60864	49519	59781
Mean	33610	56391	57739	47974	85981
Stand dev	3447	12596	7001	16140[2]	19924

[1] The value of the constant scale elasticity.

[2] Exclusive of the largest value.

* $\varepsilon_{max} = \max\left(1/(\alpha+\beta x)\right) < 1$; ε decreasing.

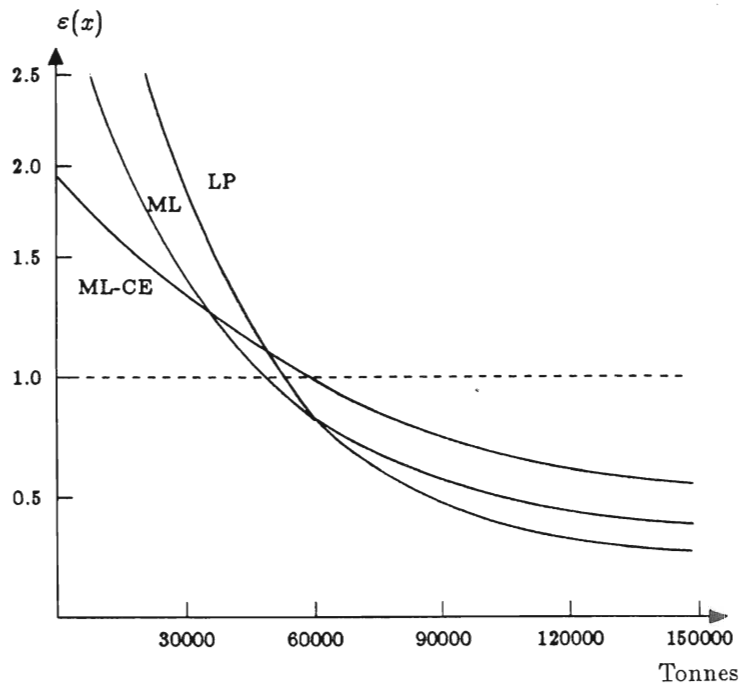
** Increasing scale economies for $x \in (0, 1163296)$, $\varepsilon \in [1.07, \infty)$.

or two medium-sized units. The ML-frontier seems to be more reluctant to bend downwards over the whole set of data close to the observed units at both tails of the size distribution, but is close to the LP-frontier for units in the medium range of output levels. This is demonstrated in Figure 7.7 for 1973, a year when the ML-close-to-the-frontier units all were of middle-range size; the same could also be said about the set of LP-frontier units.

In comparison with the LP and ML-frontiers the ML-CE-frontier appears through the data set more like an average production function and does not bend down for small and large output values.

Elasticity of scale function

Instead of comparing the individual scale parameters α and β, it is more relevant to compare the scale functions. Key differences are revealed by the technically optimal scale values set out in Table 7.2.



$$\varepsilon(x) = 1/(\alpha + \beta x)$$

Figure 7.8: The plotting of the elasticity of scale functions for 1973.

The ML-CE-results of the optimal scale output levels are systematically higher than the ML and LP-results, except for one year with a very high ML-value (and for the years with a constant scale elasticity LP-value greater than 1). The ML and LP-values are rather similar at least for five years and for three years the differences are very small. With the exception of one ML-value the estimated optimal scale output levels are systematically higher than the observed average output levels.

These differences in the technically optimal scale values for the different estimation methods can be explained by their features, as we have already discussed above. One would expect that the "average-like" ML-CE-method should result in a higher optimal scale output value than the LP method, since the latter by its very nature gives a frontier bending around the data set from above.

The scale elasticity functions are plotted in Figure 7.8. The LP scale

function slopes more steeply than the ML function and intersects the ML function from above. For the same output interval the ML-CE scale function varies less than the other two, predicting higher scale elasticity values for output levels above 30,000 tonnes of milk than the ML scale function, and predicting higher values for output levels above 45,000 tonnes than the LP-scale function. This flatter ML-CE scale function is in accordance with the observations in Figure 7.7.

Sensitivity analysis

As seen from the LP-results, there are some corner solutions for the parameters. This may be due to the fact that one unit is "dominantly" efficient. Therefore, it may be of interest to perform a sensitivity test based on removing, for each cross-section sample, that unit with the highest shadow price for the on-or-below-the-frontier constraint (4.7).[3]

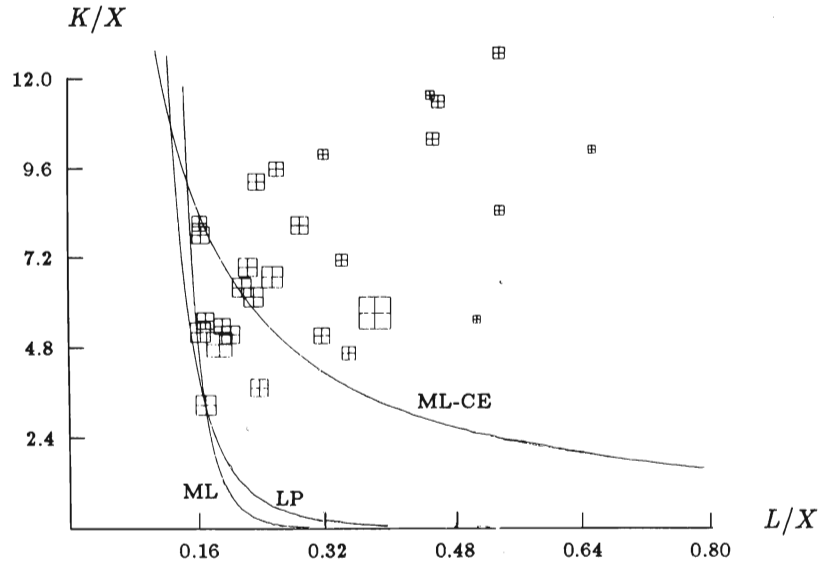
The main result (denoted by LP* in Tables 7.1 and 7.2) is that all corner solutions disappear. With respect to each parameter in Table 7.1, the change in the constant term is unsystematic, while the kernel and scale parameter values tend to vary less between the years. The greater stability of the scale parameters is reflected in the optimal scale values set out in Table 7.2. Compared with LP, ML and ML-CE, the LP* results are on the average more stable for all parameters in Table 7.1. On the other hand and in contrast with Timmer [1971] the LP* function does not move into the direction of the average function.

Efficiency frontiers

Another feature of the production functions can be illustrated in the input-coefficient space by the shape of the technically optimal scale curves or the efficiency frontiers.[4] The efficiency frontier is the locus of all points where the elasticity of scale equals 1, i.e., it is a technical relationship between inputs per unit of output for production units of optimal scale. Thus, the efficiency frontier represents the optimal scale of the frontier production function. In the input-coefficient space the frontier or ex ante production function defines the feasible set of production possibilities, while the efficiency frontier is a limit towards the origin of this set.

[3] We think this is a more appropriate way of carrying out a sensitivity test than removing all the frontier units as is done in Timmer [1971].

[4] See Section 3.3 and Førsund and Hjalmarsson [1974a].



$$(L/x)^{a_L}(K/x)^{a_K}A(e\beta/(1-\alpha))^{\alpha-1} = 1$$

Figure 7.9: The efficiency frontiers (optimal scale curves) for 1973.

The optimal scale curves for the three specifications for 1973 are plotted in the input-coefficient space as shown in Figure 7.9. The observations are represented by squares proportional to the observed output level. (The centre points are coordinate points.) Due to the imposed restrictions no observed point can be to the left of the LP and ML-efficiency frontiers, whereas 9 of the 28 observations are to the left of the ML-CE-efficiency frontier. The shape of the efficiency frontiers reflect the differences in the elasticities for that year as seen in Table 7.1 The ML-CE-variant has the lowest value for the labour elasticity and rather similar values for the ML and LP-results resulting in similar graphs for the corresponding efficiency frontiers.

With respect to the observations there is one unit quite close to the LP and ML-*efficiency frontiers*. This unit is actually on the corresponding *production frontier*. However, it does not appear to be an "outlier" in an

abnormal sense. The rest of the observations are spread fairly evenly in the region northeast of the most efficient observations. The smallest units tend to be the least efficient when efficiency is interpreted as the distance from the efficiency frontiers.

According to the ML-CE-results, the 9 observations to the left of the efficiency frontier in question are there almost exclusively due to random variation only. However, it should be noted that it is almost the same set of units that has been to the left of the ML-CE-frontier for these last three years.

Tentative interpretations of the results

When assessing the differences pointed out in previous sections one must keep in mind the different natures of the frontier estimation approaches implied by the three methods. As regards the more or less unsystematic pattern from year to year of the results of the programming methods, it is natural to expect the results to be more sensitive to "outlying" observations when "on-or-below-the-frontier constraints" such as in (4.7) are imposed. By the very nature of the programming estimation procedures we would expect some observations (or at least one) each year to be on the frontier, implying that when the set of on-the-frontier observations varies from year to year the shape of this frontier will be more affected than when, as in the case of the ML-CE method, the estimation is in fact based upon the overall movement of the set of observations.

Since only the ML-CE method yields standard errors on the parameter estimates, it is worthwhile to comment on them further. The variance of the composed variable $\ln u$ is[5]:

$$\text{var } (\ln u) = \sigma^2 + 1/(1+a)^2 \tag{7.2}$$

The impact of introducing the symmetric distribution is clearly revealed by the result that for six of the ten years included in the study over 80 per cent of the variance of the composed variable is due to the symmetric random variable. For three years the shares are above 90 percent.

The actually estimated ML-CE parameters and their standard errors are set out in Table 7.3. Table 7.4 presents the corresponding results of the "average" model. In general the asymptotic standard errors are rather small (around 10 per cent of the estimated value and even less in some cases), except for the scale and efficiency parameters.

[5] See Meeusen and van den Broeck [1977a].

Table 7.3: Results of the composed error model[1] (x in ktonnes). Asymptotic standard deviations below the estimates.

Year	Constant	Kernel elasticities		Scale parameter		Efficiency parameter
	$\frac{\ln(A \cdot 10^{-3\alpha})}{\alpha}$	a_L/α	a_K/α	$\beta \cdot 10^3/\alpha$	σ	$1+a$
1964	− 9.5060	.7957	.4440	.00197	.1627	10.8640
	(.3595)	(.0045)	(.0036)	(.00195)	(.0291)	(2.9350)
1965	− 7.8266	.6935	.3797	−.00086	.1936	18.6910
	(.1958)	(.0834)	(.0604)	(.00219)	(.0573)	(3.2921)
1966	− 9.3265	.6915	.5110	.00231	.2492	29.8410
	(2.4490)	(.1459)	(.1155)	(.00214)	(.0387)	(.7696)
1967	−12.0450	.8023	.6643	.00505	.2182	8.3821
	(.0494)	(.0013)	(.0009)	(.00117)	(.0071)	(.4324)
1968	−12.5180	.9278	.6080	.00542	.2329	8.7530
	(3.2360)	(.2911)	(.2737)	(.00572)	(.1283)	(2.4640)
1969	−13.6821	.9527	.6987	.00716	.1652	5.0490
	(.0735)	(.0011)	(.0010)	(.00093)	(.0446)	(1.3681)
1970	−10.3132	.5515	.7204	.00292	.2243	5.1636
	(.3269)	(.2204)	(.1537)	(.00320)	(.0302)	(2.2573)
1971	−15.8974	.9630	.9084	.01402	.3820	5.4489
	(1.9041)	(.2708)	(.0762)	(.00312)	(.0776)	(2.4033)
1972	−18.8510	1.3574	.8889	.02032	.4500	4.7275
	(.0398)	(.0019)	(.0016)	(.00204)	(.0921)	(.5033)
1973	−16.0860	.9737	.9290	.01510	.4535	10.2407
	(.0856)	(.2818)	(.2067)	(.00328)	(.0626)	(.0859)

[1] The underlying production function is of the form $xe^{\beta' x} = A' L^{a'_1} K^{a'_2}$. For computational reasons we have expressed the production output in ktonnes. Consequently, to transform the original production function estimates to the estimates in Table 7.1 we have to adapt the constant term and the scale parameter for this dimensional change.

Table 7.4: Results of the average production function (x in ktonnes). Asymptotic standard errors below the estimates.

Year	Constant	Kernel elasticities		Scale parameter	
	$\frac{\ln(A \cdot 10^{-3\alpha})}{\alpha}$	a_L/α	a_K/α	$\beta \cdot 10^3/\alpha$	σ
1964	− 9.562	.7980	.4389	.00193	.1879
	(2.077)	(.1703)	(.1249)	(.00462)	(.0341)
1965	− 7.854	.6916	.3789	−.00092	.2004
	(1.333)	(.1296)	(.1151)	(.00271)	(.0780)
1966	− 9.350	.6886	.5110	.00221	.2493
	(2.518)	(.1983)	(.1723)	(.00510)	(.0490)
1967	−12.127	.8039	.6597	.00500	.2520
	(2.625)	(.2277)	(.1766)	(.00514)	(.0478)
1968	−12.632	.9287	.6069	.00532	.2595
	(3.087)	(.2786)	(.2024)	(.00557)	(.0528)
1969	−13.353	.9592	.6483	.00639	.2507
	(3.119)	(.3080)	(.1885)	(.00500)	(.0524)
1970	−10.830	.5925	.7210	.00341	.2967
	(1.990)	(.2015)	(.1818)	(.00367)	(.0274)
1971	−15.923	.9464	.9065	.01351	.4213
	(2.918)	(.3099)	(.2406)	(.00491)	(.1126)
1972	−18.851	1.3574	.8889	.02032	.4500
	(3.202)	(.3590)	(.2825)	(.00496)	(.1682)
1973	−16.156	.9706	.9288	.01506	.4545
	(5.504)	(.3285)	(.2751)	(.00720)	(.1316)

Comparing results of the average function and the composed error model, we conclude — like in previous research[6] — that in most cases the asymptotic standard errors in the composed error model are of the same order of magnitude, but slightly smaller, than in the average model.

As shown by the comparison of Tables 7.3 and 7.4 the estimated parameters of the ML-CE-model have been affected very little. Except for 1969, the intercept changes only slightly in an upward direction; the same applies to the scale parameters, with the exception of 1970. As in previous implementations, the production frontier tends to be a neutral shift of the average production function.

Pooled sample results

The cross-section results above reveal considerable shifts in the production frontier from year to year. From an empirical point of view it seems more plausible that the frontier changes more gradually over time. We will investigate this possibility by pooling the data, assuming the same frontier for all years except for the constant term, which is assumed to have a time trend (i.e., Hicks-neutral technical change). The results are shown in Tables 7.5 and 7.6.

As Tables 7.5 and 7.6 reveal, the pooled results are a good representation of the average of the cross-section results shown in Tables 7.1–7.3. The standard errors of the ML-CE results are still small. In all cases the optimal scale output values are somewhat lower than the average of the cross-section results. The similarity between the ML-CE and average function results is very high. The ML-CE case yields a considerably lower rate of technical progress than the LP and ML cases, and just a little bit higher than for the average function.[7] In this sense the ML-CE function is an "average-like" function.

Technical progress, in Hicks-neutral terms, appears to have been rapid, amounting to about 6 per cent yearly in the LP-case and even higher in the ML-case. The high rate of technical progress in the ML-case is combined with a relatively low technically optimal scale output value compared with the LP-result. There is a clear tradeoff between technical progress and op-

[6] See Meeusen and van den Broeck [1977a, 1977b].

[7] This difference between the rate of technical progress in the average function and the LP-frontier has been observed earlier in Førsund and Hjalmarsson [1978a] with approximately the same figures as in this study for the rate of technical progress in the average function and the LP-frontier, respectively.

Table 7.5: Estimates of the frontier production functions and the average model. Combined time-series cross-section analysis. Estimates of the production function $x^{\alpha}e^{\beta x} = e^{\gamma t}L^{a_L}K^{a_K}$ ($t = 1$ in 1964, $t = 10$ in 1973).

Case	Constant term $\ln A$	Trend A $\gamma \cdot 10^2$	Kernel elasticities a_L	Kernel elasticities a_K	α	$\beta \cdot 10^5$	Technically optimal scale	$E(u) = \frac{1+a}{2+a}$
1 LP	−7.58	6.22	.73	.27	.19	1.52	54425	
2 ML	−2.25	10.16	.86	.12	.71	.68	42712	.66
3 LP-CE	−3.63	3.20	.56	.44	.68	.43	73800	.94
4 Average	−3.69	3.15	.54	.46	.68	.44	73703	

Table 7.6: Results of the composed error model (x in 1000 tonnes). Combined time-series cross-section analysis. Asymptotic standard deviations below the estimates.

Case	Constant term $\frac{\ln(A \cdot 10^{-3\alpha})}{\alpha}$	Technical change γ/α	Kernel elasticities a_L/α	Kernel elasticities a_K/α	$\frac{\beta \cdot 10^3}{\alpha}$	σ	Efficiency parameter $1+a$
2 ML-CE	−12.223	.04683	.8150	.6482	.00627	.2984	13.920
	(1.050)	(.00644)	(.0477)	(.0124)	(.00099)	(.0177)	
3 Average	−12.337	.04639	.8003	.6716	.00640	.3109	
	(1.076)	(.00920)	(.0934)	(.0648)	(.00218)	(.0220)	

timal scale estimates.[8] The influence of these two effects on the movement of the efficiency frontier is similar. A relatively low optimal scale together with a relatively high rate of technical progress means that the efficiency frontier *starts* relatively distant from the origin but moves relatively more rapidly towards the origin, as compared with the case with a relatively high optimal scale but with a low rate of technical progress.

Against the background of this rapid technical progress for an industry with long-lived equipment one should expect a great dispersion between the input requirements of different units, i.e., one should expect a low value of structural efficiency, $E(u)$, which is quite contrary to the ML-CE results in Table 7.5 and more in accordance with the ML-results.[9]

7.5 Frontier production functions and technical progress

Introduction

The purpose of this section is to analyse technical progress in Swedish general milk processing in terms of the frontier production function in accordance with the framework discussed in Section 4.3. We will translate the shifts in the production function, allowing for non-neutral technical change and changes in optimal scale, into a reduction in unit-costs. This unit cost reduction is split multiplicatively into parts due to neutral technical advance, factor substitution and increase in optimal scale.[10]

The analysis is based on the complete set of cross-section time-series data for 10 years, 1964–73, of the 28 individual plants.

We have utilised the homothetic production function with a variable scale elasticity analysed in Section 4.3.

$$G(x,t) = x^{\alpha-\gamma_4 t}e^{(\beta-\gamma_5 t)x} = g(v,t) = Ae^{\gamma_3 t}\prod_{i=1}^{2} v_i^{a_i-\gamma_i t} \qquad (7.3)$$

where x = output, v = vector of inputs, $G(x,t)$ is a monotonically increasing function, and $g(v,t)$ is homogeneous of degree 1 in v. Technical change is accounted for by specifying the possibility of changes in the constant

[8] See also Sato [1978].

[9] See also Chapter 3.

[10] See Section 3.6.

term A, and the kernel elasticities a_j for labour L, and capital K, and the scale function parameters α and β.

With respect to the generation of the actual data, several schemes can be envisaged. One hypothesis is that the production structure is of the putty-clay type[11] with simple Leontief (limitational) ex post functions. To simulate the actual performance of plants an efficiency term with respect to the utilisation of the inputs distributed in the interval $(0, 1]$ can be introduced multiplicatively on the r.h.s. of (7.3). We adopt this scheme and in addition assume that the plants are operated on the "efficient corners" of the isoquants. Ex post the plant managers can only choose the rate of capacity utilisation. With these assumptions concern about the "slack" in fulfilling marginal conditions with respect to inputs is not relevant. The frontier function might be regarded as a pessimistic estimate of the ex ante or planning production function. However, it is not possible at our level of aggregation to identify unique vintages. Technical change is characterised by successive improvements, while we assume discrete time with one year as the unit and fixed coefficients for each year.

In order to keep the estimation problem as simple as possible we chose here to minimise the simple sum of deviations from the frontier with respect to input utilisation after logarithmic transformation, subject to on-or-below-frontier constraints. With this specification the estimation problem is reduced to the most simple problem of solving a standard linear programming problem.

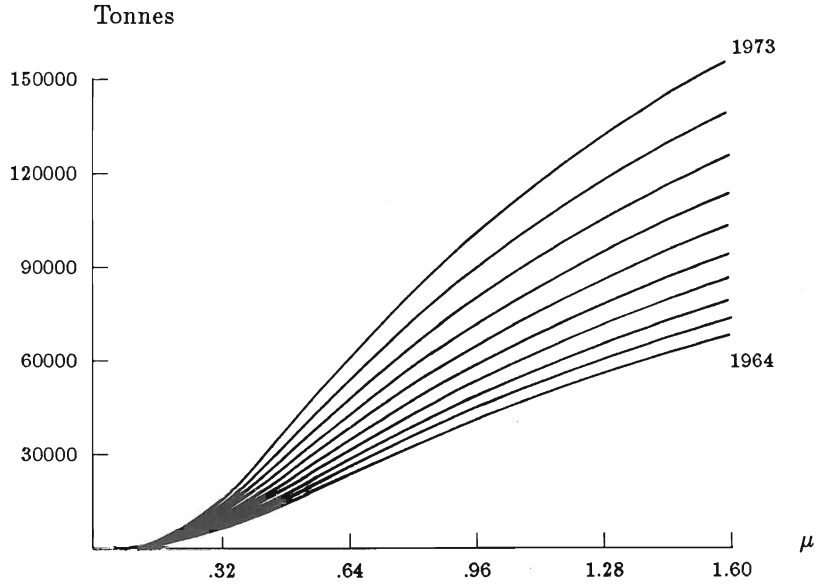
Empirical results: frontier estimates

The estimates of the parameters of the frontier production function are shown in Table 7.7 and the figures below. The different performed runs have been denoted Case 1 to Case 4. Case 1 is regarded as the main case, while the other cases represent the basis for the sensitivity analysis. In Case 2, the sensitivity of trend specifications is shown, since only Hicks-neutral technical progress is assumed. In Cases 3 and 4 another kind of sensitivity analysis is performed. In Case 3 we have excluded the largest plant from the sample and in Case 4 we have excluded the four smallest plants. The results show the sensitivity with regard to the observations.

[11] See Johansen [1972].

Table 7.7: Estimates of the frontier production function. Combined time-series cross-section analysis. Estimates of the production function $x^{\alpha-\gamma_4 t}e^{\beta-\gamma_5 t} = Ae^{\gamma_3 t}L^{a_1-\gamma_1 t}K^{a_2-\gamma_2 t}$ ($t = 1$ in 1964, $t = 10$ in 1973).

Case	Constant term $\ln A$	Trend A $\gamma_3 \cdot 10^2$	Labour elasticity $a_1 - \gamma_1 t$		Trend L $\gamma_1 \cdot 10^2$	Capital elasticity $a_2 - \gamma_2 t$		α	Trend α $\gamma_4 \cdot 10^2$	$\beta \cdot 10^5$	Trend β $\gamma_5 \cdot 10^6$	Optimal scale x for $\varepsilon = 1$	
			1964	1973		1964	1973					1964	1973
1 28 × 10	−6.02	0.	.81	.86	−.55	.19	.14	.32	.56	1.47	.73	48644	99325
2 28 × 10	−7.58	6.22	.73	.73		.27	.27	.19		1.52		53425	53223
3 27 × 10	−6.81	0.	.83	.91	−.91	.17	.09	.22	.62	2.14	1.07	38158	77818
4 24 × 10	−8.83	0.	.72	.74	−.19	.28	.26	.05	.13	2.02	1.01	49613	95284



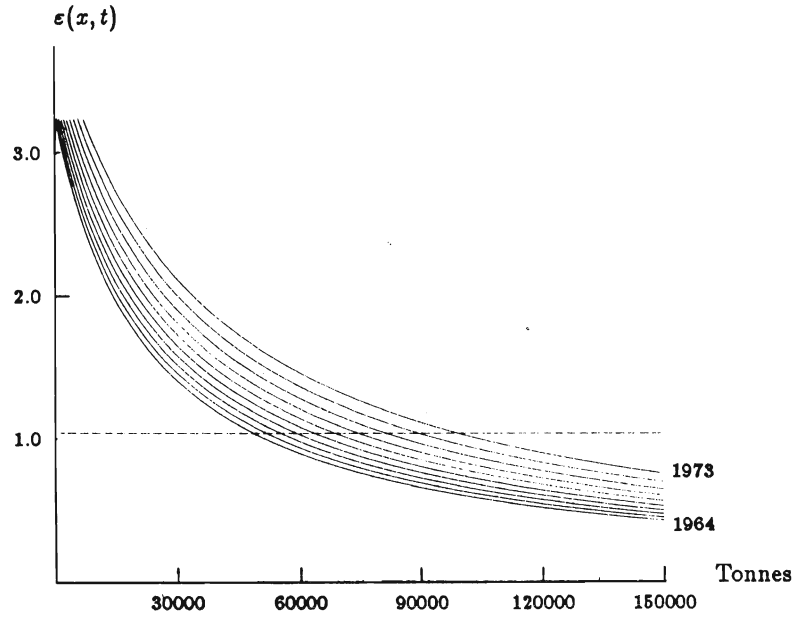
A vertical plane through the origin cutting the production function along a ray, $(\mu L^0, \mu K^0)$, $L^0 = 13000$ and $K^0 = 200000$
$x^{\alpha-\gamma_4 t}e^{(\beta-\gamma_5 t)x} = Ae^{\gamma_3 t}(\mu L^0)^{a_1-\gamma_1 t}(\mu K^0)^{a_2-\gamma_2 t}$

Figure 7.10: The change in the frontier production function through time. Combined time-series cross-section analysis.

The main result

Technical change for Case 1 is characterised by an increasing kernel elasticity of labour and a mirror image decreasing kernel elasticity of capital. For constant factor prices this implies that the units should increase the labour-capital ratio. In this sense the technical change can be characterised as capital saving. Capital-saving technical progress means in our context that the marginal productivity of labour is increasing over time.

The estimated trend in the scale elasticity function implies a considerable increase in optimal scale—about a doubling during the period. The Hicks-neutral term turned out to be on its zero lower boundary. The impact on the production surface of these changes is shown in Figure 7.10.



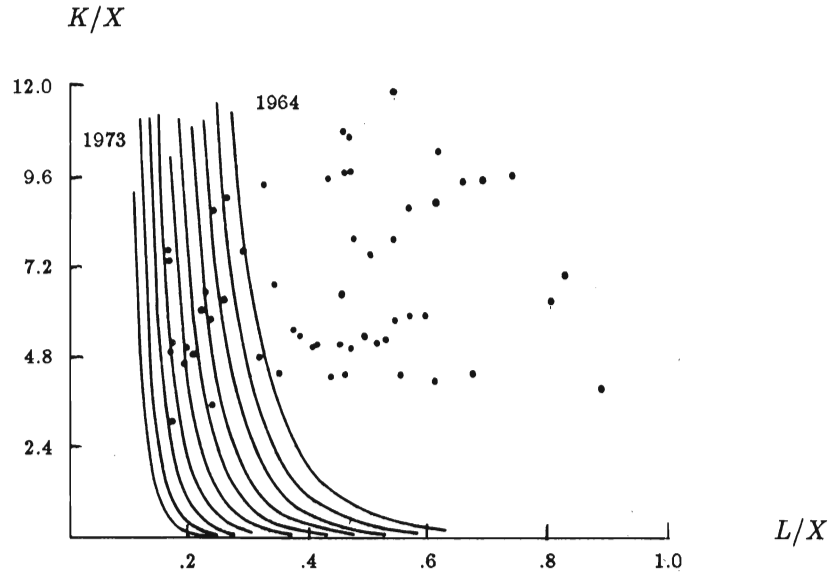
$$\varepsilon(x,t) = \frac{1}{\alpha - \gamma_4 t + (\beta - \gamma_5 t)x}$$

Figure 7.11: The plotting of the elasticity-of-scale function for all 10 years.

Cutting the production function with a vertical plane through the origin along the average factor ray, a ray corresponding to the average factor ratio, one obtains the classical text-book S-shaped graph of the production function. For this average factor ratio the development through time gives the impression of rapid technical progress due to the increase in optimal scale.

The shift in the elasticity-of-scale function can be studied in Figure 7.11, where the function is plotted for different years. The level of $\varepsilon = 1$, i.e., optimal scale is indicated. The scale elasticity shifts through time in such a way that optimal scale increases at an accelerating rate, from 6 per cent at the start to 10 per cent at the end of the period.

The smallest plant in 1964 is about 10,000 tonnes and in 1973 about 8,000 tonnes. The output of the largest plant has been within the interval



$$\left(\frac{L}{x}\right)^{a_1-\gamma_1 t}\left(\frac{K}{x}\right)^{a_2-\gamma_2 t}\cdot Ae^{\gamma_3 t}\cdot\left(\frac{e(\beta-\gamma_5 t)}{1-(\alpha-\gamma_4 t)}\right)^{\alpha-\gamma_4 t-1}=1$$

Observed units denoted by squares for 1964 and crosses for 1973.

Figure 7.12: The changes in the efficiency frontier through time.

111,000 and 141,000 tonnes in the period 1964–73, while the average output has increased from 29,000 tonnes to 39,000 tonnes, compared with the estimated optimal scale increasing from 49,000 to 99,000 tonnes.

Thus the largest unit had a scale elasticity less than 1 during the period, whereas the average output corresponds to scale elasticities considerably greater than 1.

It is obvious from Figure 7.10 that the production function is not concave over its entire domain. In Førsund [1974] it is shown that the production function with the functional specification utilised in this chapter is concave for the values of output corresponding to $\varepsilon < 1/\alpha$. In Case 1 the estimate of α is .32 in 1964 and .27 in 1973, yielding that the production function is concave for $\varepsilon < 1.77$ in 1964 and $\varepsilon < 1.92$ in 1973, which corresponds to an output of 17,583 and 33,961, respectively.

The characteristics of technical advance can also be illustrated in the input-coefficient space by the development of the efficiency frontier.[12] The development of the efficiency frontier and the observed input coefficients for 1964 and 1973 are shown in Figure 7.12. Note that for homothetic functions the shape of the efficiency frontier is identical with the shape of the isoquants.

The speed with which the efficiency frontier moves towards the origin is clearly exhibited. For instance, along the ray of the average factor ratio, the input coefficients on the 1973 frontier are about 40 per cent of the input coefficients on the 1964 efficiency frontier. It is also interesting to note that 17 of 28 units in 1973 have passed the 1964 efficiency frontier.

The increasing slope of the efficiency frontier illustrates the capital saving bias even if the trends in the kernel elasticities of labour and capital are rather small. The estimated capital saving technical progress is contrary to what one would guess a priori. Examples of labour saving techniques which have been introduced in the dairies are easy to find: Changes of milk reception from cans to tanks, self-cleaning separators and one story buildings. The observed capital-labour ratio has increased substantially for all the production units over the ten-year period. Figure 7.12 reveals that all the units have reduced their input coefficients of labour, while about half of the input coefficients of capital have increased. But the relative price increase of labour has been considerably higher than that for capital, the price indexes for the last year being 2.45 and 1.60 for labour and capital, respectively (1 for the base year).

Sensitivity analysis

It is inherent in our approach that the traditional statistical test possibilities are missing. In place of these we have performed some sensitivity tests.

In Timmer [1971] a kind of sensitivity analysis was performed by estimating the "probabilistic" frontier. This was done by discarding efficient units on the frontier from the first run and then reestimating a new frontier without the most efficient units. The purpose was to investigate the effect of the most "extreme observations". The result was that the new frontier without the "extreme" observations differed a lot from the original frontier, but was more similar (except for the constant term) to the traditional average production function for the same data-set. When assessing

[12] See Chapter 3.

frontier estimation, however, one must keep in mind that the raison d'etre of frontier function estimation is that the most efficient units should count disproportionally.

However, we are more interested in another kind of sensitivity analysis. Since there is one dominating firm in our sample we are interested in its influence on the scale properties of the production function. Incidentally, the dominating firm is only once on the frontier. The influence of the smallest plants, of which one is on the frontier, on the results is also of interest because we presume that if these plants were to be built today new and more efficient techniques might be available for the same scale of output. The Hicks-neutral case is, of course, also of interest because most earlier studies have been limited to this one case.

In Case 2 with only neutral technical progress the elasticity-of-scale function is constant and optimal scale obtains a moderate value, somewhat higher in 1964 than for Case 1, but considerably lower in 1973. On the other hand the trend in the constant term is now rather high, so neutral technical progress amounts to about 6 percent, which is a rather high value.[13] Labour elasticity is also lower and capital elasticity higher in this case. Hence with this specification, a 60 percent higher capital-labour ratio is optimal for the same relative factor prices than for Case 1 in 1964, and 130 per cent in 1973.

The objective function, the sum of slacks, increases by 3.6 per cent from Case 1 to Case 2, and is thus not negligible. In Case 1, 6 units were on the frontier, while in Case 2, 5 units were on the frontier. Moreover, in Case 1, one unit is on the frontier in 1973, the unit with the lowest input coefficient of labour. But in Case 2 no unit is on the frontier after 1971. With the flexible specification in Case 1 it pays in terms of a reduced objective function to shift the ratio between the kernel elasticities in favour of labour, such that this highly labour efficient unit appears on the frontier.

The exclusion of individual observations in Cases 3 and 4 has some influence on the results. The exclusion of the largest plant in Case 3 reduces optimal scale and increases capital-saving bias. An inspection of the data shows that the input coefficients of labour and capital have been very stable for this plant, and this has tended to reduce the capital-saving bias. The opposite is true for the four smallest plants whose input coefficients for labour, which are among the highest in the sample, have decreased relatively more than for most of the other plants. This explains the large reduction in capital-saving bias in Case 4, where all these small plants are

[13] Cf., Ringstad [1971].

excluded. In this case, however, the level and development of optimal scale is very similar to Case 1.

If small obsolete plants are included, the frontier may give a pessimistic bias over the relevant range. However, removing these units has created a much stronger bias. The small units are not replaced by observations of technologically new plants of the same scale, so really we have no control over what happens with the frontier. It turns out that the four smallest plants now in the sample are very close to the frontier, with one small unit on the frontier at the start and another at the end of the period.

The characterisation of technical change

In order to assess the importance of the various parameter changes reported in Table 7.7 we will here adopt the framework presented in Sections 3.6 and 4.7 for characterising technical advance by relative change in total unit cost, assuming cost minimisation, constant factor prices and relative change in factor ratios for constant factor prices (bias measure). The estimated technical advance measures are set out in Table 7.8 for the observed average factor ratio.

For the first two years the overall technical advance measure is $T = .92$, i.e., the average cost at the optimal scale in the second year is 92 per cent of the average cost at optimal scale in the first year, representing a decrease in the average cost of about 9 percent. Between the last two years technical advance is somewhat more rapid, about a 13 per cent decrease in average costs. Overall technical advance, T, is the product of proportional technical advance T_1, and factor bias advance T_2 In our case technical advance is due to the movement of the efficiency frontier towards the origin, the factor bias advance T_2 representing only .01 per cent of the reduction in average cost. The splitting up of the proportional advance measure T_1 reveals that the cost saving is due to the change in the optimal scale: OS increases by about 10 per cent at the start of the period and by 14 per cent at the end. The factor bias puts a brake on the cost saving along the factor ray chosen. The estimated factor bias D_{LK} implies that, for constant prices or a constant factor ratio, it is optimal to increase the labour-capital ratio by 4 per cent at the start and 5 per cent at the end of the observed period. As already pointed out this change yields practically no return in terms of cost saving.

Since we have found increasing optimal scale as the driving force behind cost saving, it is of special interest to investigate the sensitivity of

Table 7.8: The Salter measure of technical advance and its components. $K/L = 15.4$ (the average factor ratio).

Type of relative unit cost reduction measures at optimal scale		28 units		27 units		24 units	
		1964/65	1972/73	1964/65	1972/73	1964/65	1972/73
T	Overall technical advance	.9207	.8882	.9186	.8816	.9415	.9038
T_1	Proportional technical advance	.9208	.8883	.9188	.8820	.9415	.9038
OS	Change in optimal scale	.9070	.8750	.8963	.8603	.9367	.8992
B	Proportional change due to bias	1.0152	1.0152	1.0252	1.0252	1.0051	1.0051
H	Hicks-neutral advance	1	1	1	1	1	1
T_2	Factor bias advance	.9999	.9999	.9997	.9995	1.0000	1.0000
D_{LK}	Rel. change in optimal labour-capital ratio	1.0377	1.0474	1.0672	1.1111	1.0094	1.0097

the overall technical advance measure when the specification of the production function is changed with respect to the development of the parameters. Allowing a time trend in the constant term only, i.e., Case 2, the overall advance measure T becomes .94, or an average cost reduction (independent of time) of about 6 percent. This is a somewhat lower cost reduction than that obtained with the flexible specification, Case 1, but still a substantial amount for a sector characterised by small day to day improvements.

The sensitivity of the results with respect to the units included in the estimation is also shown in Table 7.8. When the largest production unit is removed the results for the overall advance measure T is about the same, and when the smallest units are removed the progress is somewhat smaller. If the small units are "obsolete" with respect to relevant ex ante designs, the inclusion of these units when estimating the frontier function leads to a positive bias in the estimated technical advance. The proportional technical advance measure T_1 follows the same pattern as the overall measure T. But the impact of the change in optimal scale, OS, is somewhat greater when the largest unit is removed, and less when the smallest units are removed. Again, if these units are obsolete in the ex ante sense their inclusion gives a positive bias to the increase in optimal scale. The removal of the largest unit adds to this bias. Although the difference between the scale elasticity functions in Case 1 and Case 4 as revealed in Table 7.7 is small, it leads to a markedly slower increase in the OS-term in Case 4: 7 per cent and 11 percent, respectively, at the start and end of the period.

In Case 3 the capital-saving bias increases markedly and the optimal labour-capital ratio increases by 7 per cent at the start and by 11 per cent at the end of the period. As already mentioned the removed unit is quite stable with respect to its input coefficients. However, this increased bias has minimal impact on the cost reduction .03 per cent and .05 percent, respectively. If the units are changed over time in accordance with the relevant ex ante function it does not matter much in cost terms whether the factor ratio is optimal or not.

For Case 4 the change with respect to the bias is the opposite. The bias has now no impact on the cost reduction, and the increase in the optimal labour-capital ratio is .9–1.0 percent. It is the change within the smallest units that gives rise to the capital-saving bias, as pointed out in the previous section. If, therefore, the smallest units are technically obsolete, the technical progress is almost neutral, but with an increasing optimal scale as the driving force.

Conclusions

When variable returns to scale were allowed to be the driving force behind technical progress, it turned out to be a fairly rapid shift in the returns to scale function.[14] The upward shift of the production frontier[15] tended to be non-neutral, increasing the kernel elasticity of labour and decreasing the kernel elasticity of capital somewhat.

The splitting up of the generalised Salter measure reveals that it is the movement of the efficiency frontier[16] along a ray towards the origin that results in the significant reductions in the average costs at optimal scale, on the order of 9–13 per cent per year. Optimal adjustment to the capital saving bias results in quite insignificant cost reductions.

The sensitivity analysis showed that the production function parameters were influenced by the discarding of *a priori* chosen units, some of which turned out to be on the frontier of the complete sample. However, the form and shift of the elasticity-of-scale function were fairly stable, leading to only small variations in the cost reduction measures.

7.6 Efficiency

The framework for measurement of productive efficiency discussed in Section 3.4 is here applied to the Swedish dairy industry.

Structural efficiency

Let us first look at the aggregated picture of the industry. The estimates of structural efficiency are presented in Table 7.9 below.

The interpretation of the S_1 measure is the relative reduction in the amount of inputs needed to produce the observed industry output with frontier function technology having the observed factor proportions. Thus, the table shows that the same output in the different years could have been produced by 59–70 per cent of the observed amounts used.

The S_2 measure shows the ratio between the observed output and the output obtained for the observed amount of inputs by using frontier function technology. The table reveals that observed output is between

[14] See figure 7.11.

[15] See figure 7.10.

[16] See figure 7.12.

Table 7.9: Estimates of structural efficiency. (Definitions are found in Section 3.4.)

Year	S_0	S_1	S_2	S_3	S_4	S_5
1964	.7826	.7006	.6488	.6469	.9234	.9971
1965	.7465	.6941	.6337	.6305	.9084	.9950
1966	.7190	.6327	.5756	.5755	.9096	.9998
1967	.7018	.6264	.5622	.5619	.8970	.9995
1968	.6662	.6016	.5397	.5397	.8971	1.0000
1969	.6386	.5907	.5186	.5186	.8779	1.0000
1970	.6183	.5660	.4827	.4826	.8527	.9998
1971	.6561	.6004	.5020	.4994	.8318	.9848
1972	.6687	.6259	.5113	.5031	.8036	.9838
1973	.6475	.5928	.4715	.4658	.7858	.9879

S_0 is the weighted sum of efficiency measures.

S_1 is the distance of the average plant to the frontier function for given output. (Corresponds to E_1.)

S_2 is the distance of the average plant to the frontier function for given amounts of inputs. (Corresponds to E_2.)

S_3 is the distance of the average plant to the efficienct frontier. (Corresponds to E_3.)

$S_4 = S_3/S_1$ is the pure scale efficiency. (Corresponds to E_4.)

$S_5 = S_3/S_2$ is the pure scale efficiency. (Corresponds to E_5.)

47 per cent and 65 per cent of potential output if the inputs are employed in units with frontier production technology.

The S_3–S_5 measures show the relative reduction in input coefficients by producing at optimal scale on the frontier function with the observed factor proportions. Thus for example the table shows for S_3 that at optimal scale on the frontier production function the potential input coefficients are 47–65 per cent of the observed input coefficients.

The most remarkable result is the high level of structural inefficiency captured by all the four measures S_0–S_3. Moreover, there seems to be a clearly decreasing trend in the values of structural efficiency contrary to what most commentators on productivity differences seem to assume. Thus the distance between average performance and best practice has increased during the period. This result is confirmed in a related study, which examined the development of the distance between the frontier production function and the average production *function*.[17]

Even if the development of the efficiency measures S_0–S_3 is the same, the levels for each year differ rather a great deal. For all years, $S_0 > S_1 > S_2 > S_3$. However, the difference between S_2 and S_3 is relatively small. This means, as S_5 shows, that if the average plant is moved to the efficiency frontier in the vertical direction rather little is to be gained by moving it to the optimal scale. This stems from the fact that the average observed amounts of inputs are about the same as required at optimal scale for the first year and have developed in the same way as the amounts of inputs required at optimal scale.

On the other hand, if the average plant is moved to the frontier in the horizontal direction there still remains some pure scale inefficiency which increases rather considerably from .92 to .79 during the period. Thus most units become too small when they are moved horizontally to the frontier, a tendency which is strengthened during the period. While optimal scale has increased from about 49,000 tonnes in 1964 to 99,000 tonnes in 1973, the average output has only increased from 29,000 tonnes to 39,000 tonnes.

The low level of structural efficiency has been confirmed for one year in an earlier study by Carlsson [1972], who estimated S_0 for 26 Swedish industries in 1968 using a Cobb-Douglas frontier production function. His estimates of S_0 for the whole dairy industry in this year was 0.6184, not too far from our own estimate for that year. Moreover, it turned out that the dairy industry showed the second-lowest degree of structural efficiency of

[17] See Førsund and Hjalmarsson [1978a].

the 26 industries. What is then the reason for this high degree of structural inefficiency?

Carlsson [1972] tries to explain the differences in efficiency between industries by differences in competitive pressure and finds that protection seems to breed inefficiency. Of course, this can be one part of the explanation of efficiency differences. However, if a putty-clay production structure and embodied technical progress are empirically relevant, which seems to be the case in most manufacturing industries, normally there will be differences between production units within an industry. As pointed out in a comment on Carlsson's result, the more rapid the technical progress the less efficient the industry may appear in an analysis based on cross-section data, depending on what happens to investment and the rate of scrapping.[18] Thus, if a faster rate of technical progress increases the differences in efficiency between the best practice plants and the industry average for a given rate of industry output expansion, one might just as well state that technical progress breeds inefficiency.

The differences in efficiency can be perfectly efficient from an economic point of view, as shown in Johansen [1972] and Chapter 2 above. Important explanatory factors of industry structure at a point in time are then the forms of the establishment *ex ante* production functions within the industry, the rate of embodied technical progress and the expansion rate of the industry output.

A main characteristic of the technological structure of dairy plants is that there are different substitution possibilities before and after investments in new production techniques, i.e., one must distinguish between *ex ante* and *ex post* production possibilities. A putty-clay structure, embodied technical progress and economies of scale in plant construction give rise to different vintages of capital.

It is not possible, however, at our level of aggregation to identify unique vintages. Technical change is characterised by successive improvements of different parts of the dairies as, for example, changes of milk reception from cans to tanks and introduction of self-cleaning separators.

In Section 7.5 it is shown that technical progress has been rapid during the period. In fact, average cost at optimal scale decreased progressively from about 9 per cent per year in the beginning of the period to about 13 per cent at the end of the period.

Thus one reason, and probably the most important one, for the large and increasing differences between best-practice technology and average

[18] See Førsund and Hjalmarsson [1974b].

performance must be the underlying technological structure in combination with a rapid technical progress. Further aspects of the efficiency differences will be discussed below.

All plants included in this study have survived the entire period. During that time a number of dairies have been closed in Sweden. Thus, the development of structural efficiency for all plants may have been different from the set utilised here.

Technical efficiency and scale efficiency

The estimates of the individual measures of technical efficiency and scale efficiency are presented in Figures 7.13–7.16 below for the three different years, 1964, 1968 and 1973. In the figures the units are arranged in increasing order of their efficiency values. Each rectangle or step in the diagrams represents an individual unit. Efficiency is measured along the ordinate axis and the percentage share of (cumulative) output along the abscissa axis.

In these figures both the range and shape of the efficiency distributions are illustrated. At the same time we can observe the positions of the small and large units.

Let us first look at Figures 7.13–7.15 where the measures are shown separately. The interpretation of the measures are shown in a few examples.

In 1964 the least efficient unit according to E_1 produced about 3 per cent of total industrial output and had an efficiency value E_1 of about .50. This means that the same output could have been produced by 50 per cent of the observed amount of input when utilising best-practice technology.

The least efficient unit according to E_2 produced about 3 per cent of total output and had an efficiency value for E_2 of about .46, which means that the observed production was only 46 per cent of the output that could have been obtained had the same amount of inputs been employed in the frontier function.

Let us also look at scale efficiency E_3 in 1973. The least efficient unit, with E_3 of about .20, produced only 1 per cent of total output. If this unit had employed frontier function technology at optimal scale, the level of the potential input coefficients would have been only 20 per cent of that actually observed. The most efficient unit that year, with E_3 of about .76, produced approximately 5 per cent of total output. The level of its potential input coefficients was then 76 per cent of the actually observed if the observed amount of inputs had been employed at optimal scale in the frontier production function.

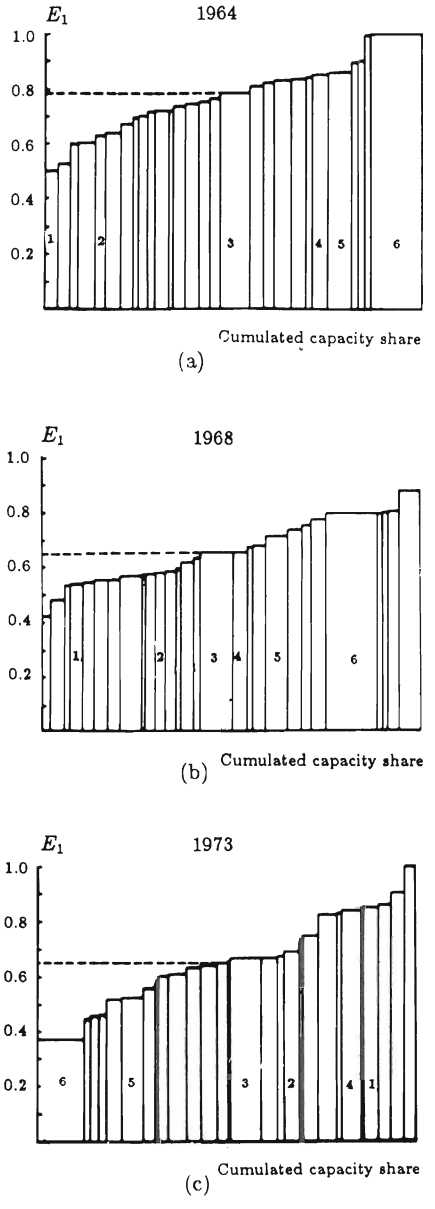


Figure 7.13: E_1 measure of technical efficiency for selected years.

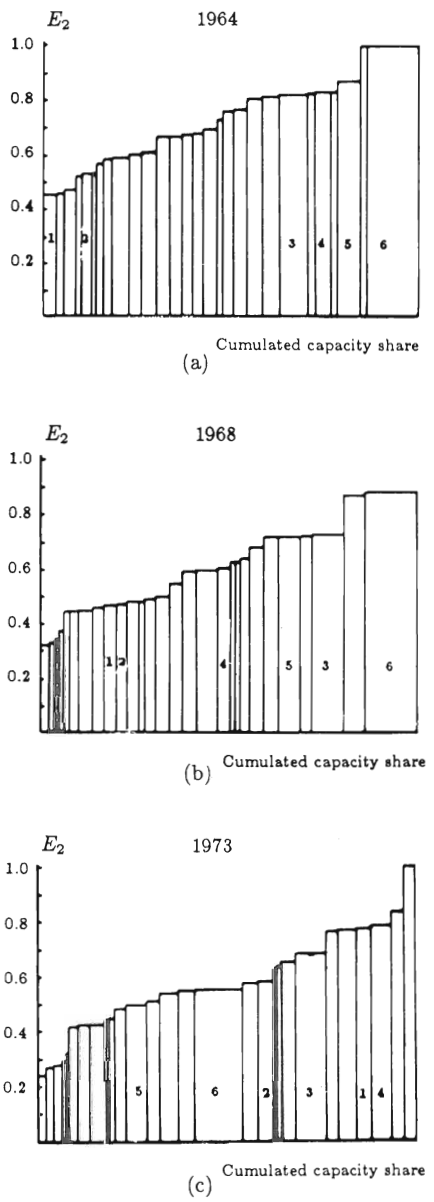
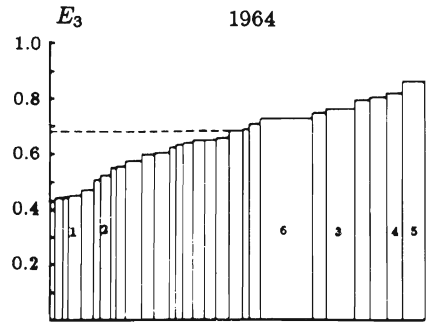
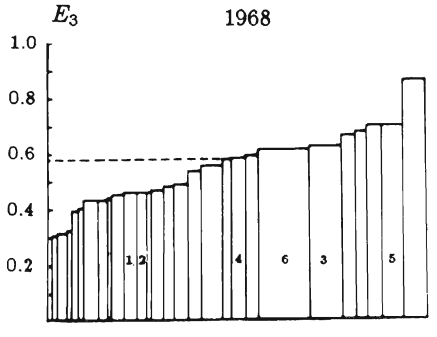


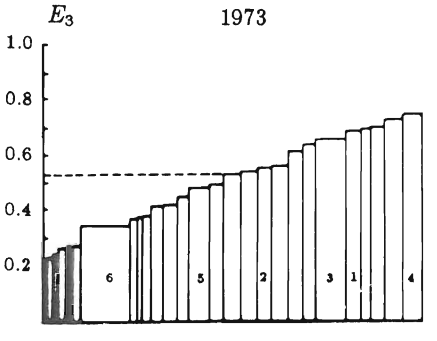
Figure 7.14: E_2 measure of technical efficiency for selected years.



(a) Cumulated capacity share



(b) Cumulated capacity share



(c) Cumulated capacity share

Figure 7.15: E_3 measure of technical efficiency for selected years.

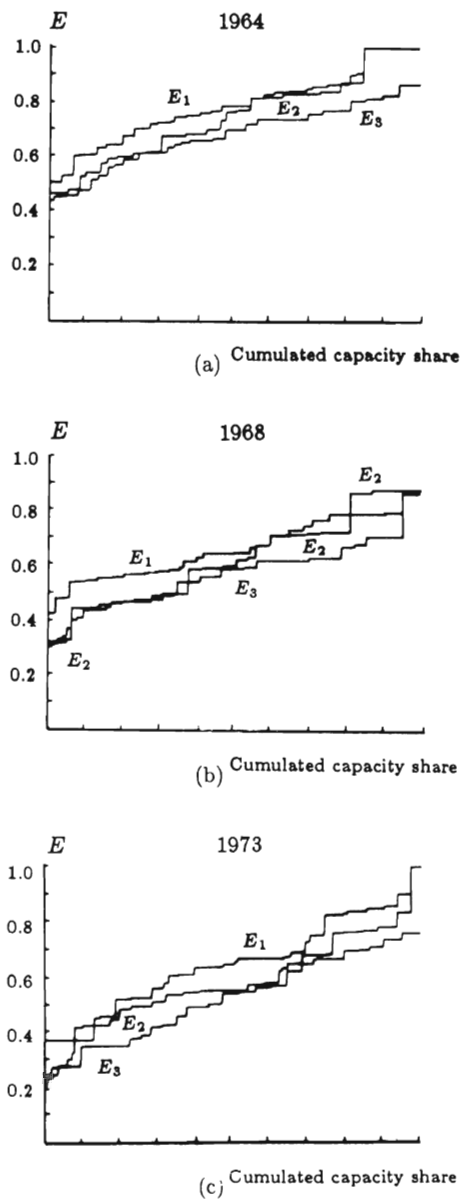


Figure 7.16: E_1, E_2 and E_3 measure of technical efficiency for selected years.

As the figures reveal, there is a large variation in efficiency between the units for all the years. The most striking example here is the E_2-values for 1973 when the most efficient unit was on the frontier ($E_2 = 1$) and the most inefficient unit had a value of $E_2 = .24$. Moreover, the range increased during the period, consistent with the development of the measures of structural efficiency.

The shape of the distributions also changed during the period. Looking at the figures from left to right, efficiency decreases rather continuously in 1964, but in 1973 the efficiency distribution becomes more irregular except for scale efficiency, which has a very regular shape during the whole period.

With respect to the position of the small and large units in the efficiency distributions, there is no clear relationship between size and technical efficiency. In 1963 the largest plant and a very small one were on the production frontier, while in 1973 it was a plant of medium size.

Six units are identified in Figures 7.13–7.15: two units, nos. 5 and 6, with declining efficiency over time, two units, nos. 1 and 2, with increasing efficiency, one unit, no. 3, with a mean value of input saving efficiency and one unit, no. 4, with the overall best performance as regards scale efficiency.

The development of the largest plant is interesting. In 1964 this plant was on the frontier, i.e. $E_1 = E_2 = 1$. In 1968 this plant was still rather efficient but in the last year its efficiency was reduced dramatically as measured by E_1 but not so much by E_2. A closer look at the data shows that the input coefficients of labour and capital were fairly constant for this unit during the period, while the input coefficients for labour decreased for most other units and were approximately constant for the capital coefficients. Hence, the productivity of this unit has been fairly constant, while at the same time the frontier has moved upwards.

Since the frontier is estimated by LP-techniques the number of units on the frontier are at most equal to the number of estimated parameters, five in this case. The frontier is also usually built up of plants of different sizes, one large, one small and a few medium sized. A very small plant with high input coefficients of both labour and capital can be on the frontier because that plant is the most efficient of that size.

The differences in the ranking of the units according to E_1 (constant output) and E_2 (constant input) are also clearly demonstrated in Figures 7.13 and 7.14. Particularly for the largest unit in 1973 the difference is striking. According to E_1 this unit was most inefficient, according to E_2 it had about medium efficiency. Therefore, it is not a matter of indifference which measure is utilised when talking about efficiency for individual plants.

For scale efficiency there is a clear tendency for the large units to show high values. An exception is the largest unit in 1973 which has a rather low value of scale efficiency.

A further comprehensive view of the development of the efficiency distributions is obtained in Figure 7.16, where all the three measures of efficiency E_1, E_2 and E_3 are plotted simultaneously as a step function, i.e., the top portions of the histograms are plotted in the same figure as an alternative to the histogram. The step diagrams give a good picture of the dispersion in the different measures and the ranking of the units according to the different measures. The total dispersion for all measures is somewhat reduced by the changes in the ranking between the different measures, while at the same time the range increases through time.

An alternative to the measures of structural efficiency above is to look at the efficiency value of that unit which covers the 50 per cent accumulated capacity point on the abscissa axis. These values are indicated by dotted lines in Figures 7.13–7.15. This median capacity value of E_1, which is very similar to the value of S_0, has decreased from .79 in 1964 to .65 in 1973. This is about the same percentage decrease as in the S_0 and S_1 measures, about 20 per cent. The median capacity value of E_2 has decreased from .77 in 1964 to .55 in 1973, which is about the same percentage decrease as in the S_2 measure (about 40 per cent) but on a higher level. The median capacity value of E_3 has decreased from .69 to .54 (28 per cent), which is a smaller reduction than that for S_3 (38 per cent).

Let us examine more thoroughly the rankings between different years of the individual units in the efficiency distributions. We are interested in investigating whether there have been any dramatic changes in the rankings of the units during the period. Therefore, we have calculated Spearman's rank correlation coefficient between consecutive years and between 1964 and 1973, together with Kendall's coefficient of concordance, denoted by W, for the whole period. The results are shown in Table 7.10 below.

The table reveals a high correlation of efficiency rankings between successive years, and highest for scale efficiency. Usually the correlation coefficient is in the interval between .80 and .95. There has not been any dramatic changes in the efficiency rankings between a pair of years and scale efficiency has been quite stable. The value of the coefficient of concordance is rather high, but somewhat lower than the correlation coefficients for successive years, also indicating high stability in the rankings.

On the other hand, there has been a gradual change in the rankings during the period, relatively small for scale efficiency but relatively large for technical efficiency. The correlation coefficient between the start and end years

Table 7.10: Spearman's rank correlation coefficient between different years and Kendall's coefficients of concordance, W.

Years	E_1	E_2	E_3
1964/65	.8544	.7969	.8402
1965/66	.7614	.8199	.9201
1966/67	.8856	.9595	.9625
1967/68	.8681	.8380	.8730
1968/69	.8027	.8210	.9373
1969/70	.7756	.7367	.7983
1970/71	.9146	.8544	.9086
1971/72	.8643	.9135	.9351
1972/73	.8593	.9245	.9688
1964/73	−.0282	.1073	.4072
W	.5429	.6011	.7003

1964 and 1973 even shows a negative sign for E_1. An example here is the largest unit which was on the frontier in 1964 but had the lowest E_1-value in 1973. The lower values of the coefficient of concordance, in comparison with the correlation coefficients for successive years, also indicate this gradual change of the rankings.

We have also confronted the dairy experts of the Swedish Dairy Federation with our empirical results and discussed the reasons for differences in efficiency between the units.

We received a confirmation that our results regarding the most and least efficient plants were reasonable. Some differences in efficiency were explained by the modernity of equipment, while others were explained by more or less skillful managements. With some simplification the small best-practice plants seemed to have good management, while large efficient plants also had modern equipment on top of good management.

When planning a new dairy it is not only the optimal scale concept estimated here that must be taken into consideration but also the collection and distribution costs and the existing structure of dairies inside a certain region.

7.7 Concluding remarks

Several new measures of efficiency have been applied to the Swedish milk processing industry. The development of the industrial structure has been studied by the change in the efficiency distributions for the individual plants through time and the aggregate performance of the sector has been studied by examining the development of the different measures of structural efficiency. The most remarkable result is the rather long distance between best-practice and average performance measured by different measures of structural efficiency. Moreover, this distance shows an increasing trend during the period. These results are explained by rapid technical progress in combination with an underlying putty-clay technological structure.

The distribution of the individual measures of technical efficiency and scale efficiency reveals a large variation in efficiency between the units for all years. Some of these differences in efficiency can be explained by the modernity of equipment and others by differences in management capability.

8

The Swedish Cement Industry

8.1 Introduction

Due to the rising fuel prices during the 1970s, the cement industry has drawn a great deal of attention as a very fuel intensive consumer.[1] Articles examining the technology of the cement industry have appeared in the economic literature. These studies, however, have mainly concentrated on the economies of scale in cement production, yielding estimates of minimum efficient scale at the *plant* level on the basis of engineering information, or statistical data from plants in operation.[2] Thus, these studies provide some insight into the scale properties of the ex ante production function for cement plants.

In this chapter we have applied the short-run function approach to an empirical analysis of technical progress and structural change in the Swedish cement industry during a twenty-five-year period 1955–79. The analysis is based on micro data for individual kilns.

8.2 Data

The cement manufacturing process

The raw material for cement production consists mainly of limestone which is crushed and then ground into a fine powder. In the dry cement manufacturing process, the powder is fed directly into a kiln where it is calcined (burned) to form clinker. In the wet process, water is added to form a

---

[1] See, e.g., Srinivasan and Fry [1981].

[2] See, e.g., McBride [1981] and Norman [1979]. An exception is Sterner [1985] who applied the short-run function in a study of the Mexican cement industry.

slurry that is then fed into the kiln. The basic principle of the semi-dry process is to use the exhaust gases from the kiln for drying and preheating the raw materials before inserting them into the kiln. Thus, the main advantage of the semi-dry process is energy saving.

The kiln is essentially a huge cylindrical steel rotary tube lined with firebrick. The raw material (either slurry or dry) is fed into the upper end. At the lower end is an intensely hot flame which provides a temperature zone of about 1500° C from the precisely controlled burning of coal, oil or natural gas under forced draft conditions. After the clinker is cooled, it is ground with 4–6% gypsum into cement.

The data set

The micro units in this study are the individual kilns of the Swedish cement industry. Cement production is usually studied at the plant level. Since the putty-clay assumptions are crucial to our approach, the kiln is the most suitable unit. The kiln is the largest and most expensive piece of equipment in the cement plant. It is the only consumer of fuel and responsible for two-thirds of the total energy consumption of the plant.

The data cover a time period between 1955 and 1979, but since our purpose is to study the long-run development of the short-run industry production function we have chosen to illustrate the results for the typical years 1955, 1960, 1965, 1970, 1974 and 1979. We have obtained all data directly from the only existing Swedish producer. The data comprises energy, labour input, capacity and actual output. Since the raw material input is strictly proportional to output, independent of vintage and size, this input is not included explicitly.

Energy consumption is measured in calories and relates to the direct use of energy for drying, heating and burning (calcining) the cement in the kiln. When different types of energy have been used we have aggregated to one energy measure based upon the raw energy content of the different energy types (primarily oil and coal). Burning coal means a small decrease in energy efficiency, i.e., that for the same amount of output, up to 5 per cent more raw energy is required from coal than from oil.

While energy consumption is kiln specific with fixed-input coefficients in the short run, labour input is not. Labour input is determined by the aggregate kiln capacity for each plant. Sticking to the kiln as the micro unit, it is then a natural assumption to allocate labour in proportions to the production of each kiln.

Table 8.1: The Swedish cement industry, 1955–79.

Year	Capacity Ktonnes	Output Ktonnes	Per cent capacity utilisation	Number of kilns
1955	2507	2502	100	18
1960	2962	2797	94	20
1965	3744	3846	103	23
1970	4967	3968	80	25
1974	4579	3738	82	20
1979	3561	2099	59	9

Since our purpose is to study the *long-run* structural change in the use of energy and labour, this procedure should yield a relevant picture of substitution and productivity changes, even if the short-run function for individual years must be regarded as an approximation of the actual production possibilities.

Capacity and output of the individual kilns are measured in tonnes of cement. According to the industry practice, annual capacity is defined as maximum daily capacity during 310 days. The industry capacity, annual output in ktonnes, percentage capacity utilisation and the number of kilns operated during the selected years are presented in Table 8.1. Note that it is possible to produce more than capacity if the number of idle days is lower than expected.

The relatively low degree of capacity utilisation in the seventies even during boom years, is due to the sharp decrease in building activity experienced in Sweden at that time. This also explains the decrease in output between 1970 and 1979. The industry still maintains old kilns as reserve capacity for peak periods.

In 1955 the whole capacity was made up of wet processes, except for one semi-dry kiln, but no wet kiln has been installed since 1967. In 1974 five kilns were dry, two semi-dry and thirteen wet. In 1979 only three wet kilns remained.[3]

[3] For a thorough description of the Swedish cement industry and its development, see Carlsson [1978].

Table 8.2: Factor price developments between 1955 and 1979 (Index 1955 = 100).

Year	Labour	Energy	Relative price
1955	100	100	1.00
1960	142	110	1.29
1965	213	95	2.24
1970	294	84	3.50
1974	510	364	1.40
1979	963	540	1.78

In this study the time unit is one year. There is empirical evidence of a more or less continuous input-saving progress, which can be considered as a certain amount of disembodied technical change. Alternatively, these input savings might be explained by capital substitution in the form of small scale investments that have been added to the basic kiln structure.

The input coefficients do, to a certain degree, depend on the rate of capacity utilisation. Both coefficients tend to increase with decreasing rate of utilisation. Due to our method of estimating the coefficients from current observations this especially affects energy coefficients for kilns with a very low rate of capacity utilisation. Stops and restarts have a negative effect on energy efficiency. Since slump years are avoided, labour hoarding should not affect the labour coefficients unduly. These qualifications underline the fact that the assumption of fixed coefficients within each year must be looked upon as a convenient approximation.

The relative prices between labour and energy have changed considerably during the period. In Table 8.2 we have calculated the factor price development on the basis of the actual costs for the cement industry. The sharp rise in energy prices in 1974 has to some degree been mitigated by an increase in coal burning, from 13 per cent of the thermal energy input in 1974 to 34 per cent in 1979.

8.3 Structural description

Labour input-coefficient distributions

Since we could not allocate labour input on the individual kilns in any exact way, all kilns belonging to the same plant have the same labour input coefficient. Figure 8.1 illustrates the distribution of input coefficients and the development of labour productivity between 1955 and 1979.

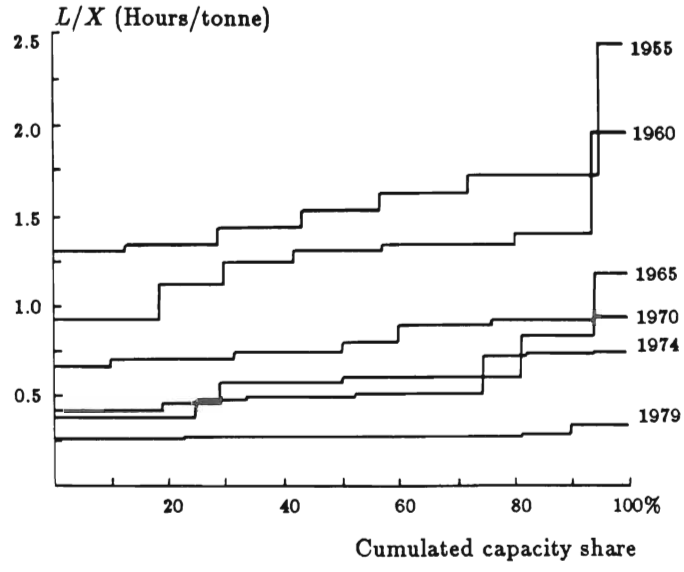


Figure 8.1: The development of the labour input coefficient distribution for selected years.

The main tendency is the gradual development towards a very flat distribution in 1979, with all units on about the same productivity level. Between 1955 and 1965 there was a rather rapid increase in labour productivity. The rate, however, slowed down between 1965 and 1974, only to increase again between 1974 and 1979. This development emerges from a gradual increase in the degree of mechanisation and automation. It is not easy to

distinguish between embodied and disembodied labour saving in the data, but on the evidence of the changes for the same kilns over the years, disembodied change could be of about the same magnitude as embodied in new kiln structures.

Energy input-coefficient distributions

The development of the input coefficients for energy is shown in Figure 8.2. The development through time is due to both disembodied and embodied technical progress, with embodied progress being the most important. Between 1955 and 1960 the entire distribution shifts downwards with the same kilns in use except for two new ones. Except for one semi-dry kiln in 1955 and two in 1960 all kilns were wet. The semi-dry kilns have the lowest energy input coefficients.

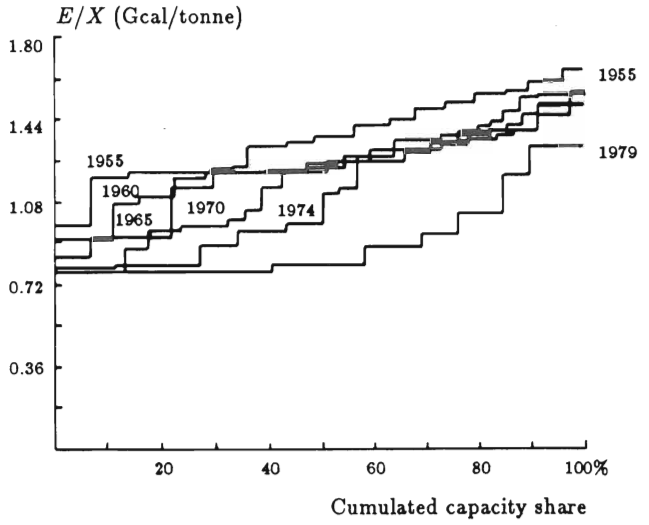


Figure 8.2: The development of the energy input-coefficient distribution for selected years.

After this shift the potential for further disembodied energy-saving technical progress seems to have been exhausted. This is illustrated by the upper 50 per cent of the capacity in Figure 8.2.

Between 1960 and 1974 the introduction of new dry kilns shifts the input coefficients of the lower 50 per cent of the capacity downwards. The

size of the kilns also increases. During this period nine new kilns, all dry except one, were installed. Seven wet kilns were closed down.

In 1979 the distribution is dominated by one large dry kiln, covering about 40 per cent of the capacity while the two largest kilns together cover 60 per cent of the capacity. In that year the best-practice technology was decisive for the shape of the distribution. Between 1974 and 1979 one new kiln was started, while eleven old kilns were closed down.

The development of the best-practice input coefficients for energy between 1970 and 1979 indicates that the potential for embodied energy-saving technical progress within the same basic technology is exhausted. The new large kiln is just slightly more energy efficient than the best kilns from 1970 and 1974.

Capacity distributions

The capacity distributions in 1955, 1974 and 1979 are shown in Figure 8.3. The capacity distribution moved considerably between 1955 and 1974, and somewhat further between 1974 and 1979, particularly in the labour-saving direction, but also in the energy-saving direction for the industry as a whole. From 1955 to 1979 the average value of the input coefficient for labour has decreased by 67 per cent as opposed to 17 per cent for energy.

While energy-input coefficients are largely embodied in the kilns, labour is not. Decreasing labour input coefficients partly reflect the increases in the size of the kilns (a larger unit does not require more labour than a smaller one), partly rationalisation in other parts of the plant. The shape of the distribution has changed somewhat due to the bulk of new dry kiln capacity and in 1979 in particular it was highly concentrated for labour-input coefficients. Except for the largest wet kiln in 1974, the largest units have also been the most efficient. The dry process makes it possible to exploit economies of scale, resulting in a labour-saving technical progress.

The fact that the largest wet kiln in 1974 had the worst energy productivity is interesting. The chemical composition of the limestone prevented utilisation of the dry process for this plant site. To exploit economies of scale a large wet kiln was installed in 1967, but problems arose and this kiln is now closed.

On the basis of actually paid average input prices we have constructed factor price ratio lines. In Figure 8.3 the factor price ratio lines are drawn through the "marginal" kilns of 1955, 1974 and 1979. Thus, we have started from actual output and calculated the cost-minimising sequence of kilns up

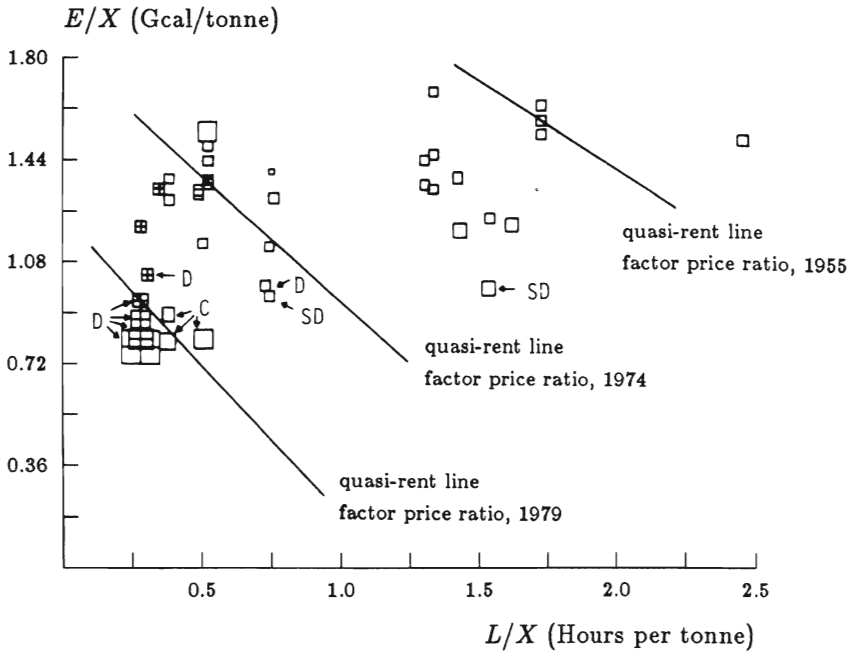


Figure 8.3: The capacity distributions in 1955 (empty squares), 1974 (empty squares) and 1979 (cross squares).

to this output. The last kiln in this sequence is the "marginal" kiln. Two kilns were above this line in 1955, five in 1974, and five in 1979.

Actually, all kilns were in use during these years but with varying degrees of capacity utilisation. This might be an indication of imperfect optimisation. However, one should take into consideration that a full optimisation of the cement industry must include the transport costs between the various plants and the consumers.

8.4 The short-run industry production function and technical change

Region of substitution

The region of substitution and isoquant map of the short-run industry production function is presented in Figure 8.4 at five-year intervals.

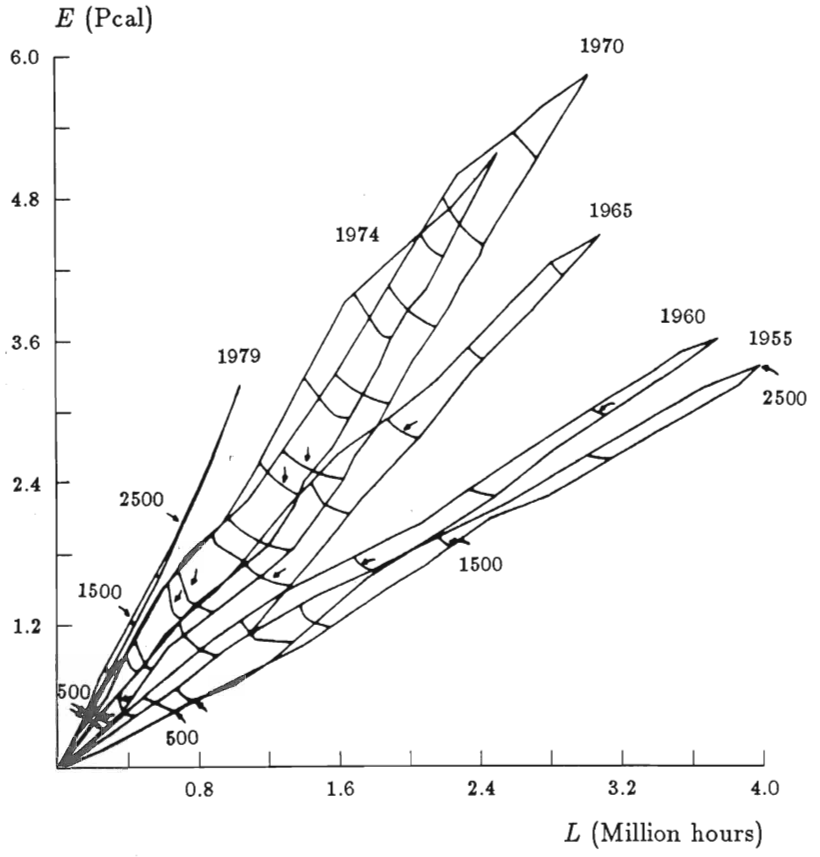


Figure 8.4: The development of the short-run industry production function between 1955 and 1979.

Comparing different years for the same isoquant level, the region of substitution is rather narrow for 1955, 1960 and 1965 and increases considerably between 1965 and 1970 when the dry process was introduced and capacity increased. An indication of this is that for the isoquant level of 2000 ktonnes the reduction in labour input (by moving from the starting point to the end point of the isoquant) was about 20 per cent in 1970–74, compared to only about 3 per cent in 1955–60. For energy reduction the values were below 3 per cent in 1955–60, as compared to about 10 per cent in 1970–74. Due to the extremely small differences in labour coefficients there was very little scope for substitution in 1979 and hence the substitution region is extremely narrow that year.

The development of the short-run function is determined by investments in new capacity and scrapping. The investment decision is based on the expected future development of input prices, ex ante technology and demand. All these factors influence the timing, factor proportions and the scale of investments. According to earlier studies there are considerable scale economies for both labour and capital in the ex ante production function while all other inputs are proportional to output.[4]

Against this background a steady shift of the substitution region towards the energy axis should be expected and is due to the simultaneous influence of the development of relative prices as shown in Table 8.2, the scale properties of the ex ante function and the shift in technology from the wet to the dry process. It is particularly important to note the reduction in labour-input coefficients due to increased scale of new kilns.

The development of the substitution region has been most rapid between 1960 and 1970 parallel to the very rapid increase in the relative price of labour. During this period the average factor ratio between energy and labour doubled,[5] and capacity increased by 68 per cent. Four relatively large, energy-economised, dry kilns were installed together with three wet kilns, while two rather energy consuming wet kilns were closed.

Demand regions

The region of substitution can also be studied partially in one dimension for each input by the demand regions, i.e., the region of feasible input utilisation for each input is presented separately as a function of relative

[4] See, e.g., McBride [1981] and Norman [1979].

[5] See Table 8.5 below.

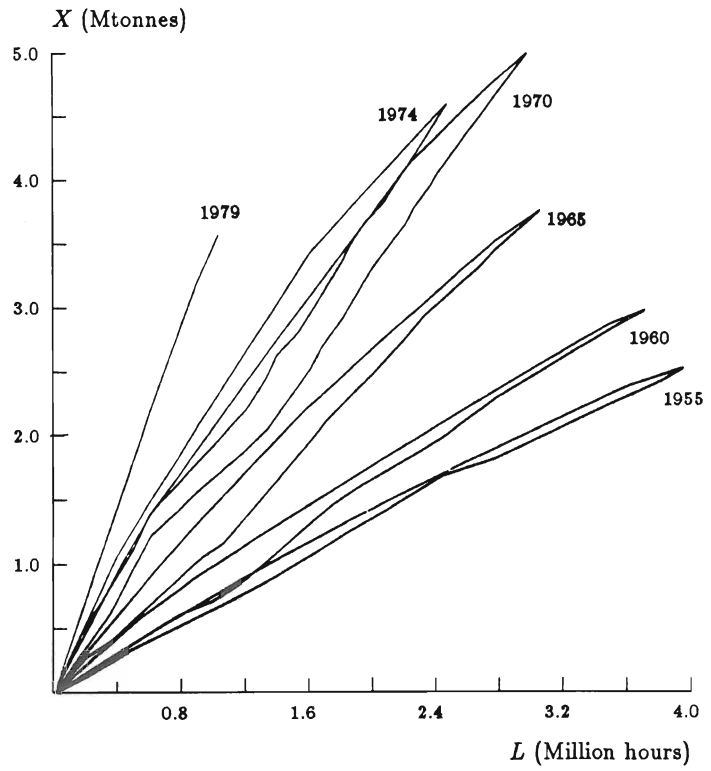


Figure 8.5: The development of the demand regions for labour between 1955 and 1979.

prices and capacity utilisation. The demand regions for labour and energy between 1955 and 1979 are shown in Figures 8.5 and 8.6.

The "curvature" or the "slope" of the demand region indicates the degree of diminishing returns for each input and the dispersion of the units in the capacity distribution diagram. The curves also confirm the impression from the partial input-coefficient distributions in Figures 8.1 and 8.2 that in 1974 the dispersion of energy-input coefficients was larger than the dispersion of labour-input coefficients.

Another point to note here is that the demand regions express the input-output coefficients for the industry as functions of capacity utilisation and relative prices. Since all isoclines lie inside the substitution region, it is

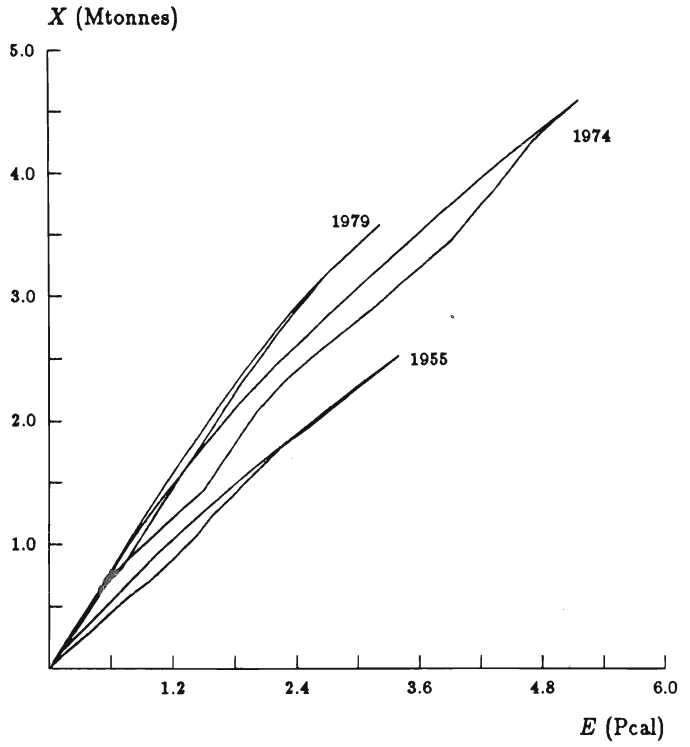


Figure 8.6: The development of the demand regions for energy between 1955 and 1979.

also possible to study the width of the substitution region and the demand regions for the interval of "relevant" factor prices.

To further elucidate the properties of factor demand functions, demand elasticities are plotted in Figures 8.7 and 8.8 for the years 1974 and 1979. There is a great difference in the shape and width of the demand regions for these years. For the narrow region of 1979 the demand elasticities are 1 at the start, increasing for output levels close to full capacity utilisation. The greater width of the 1974-region is reflected in jumps in the demand elasticities to values lower than 1.

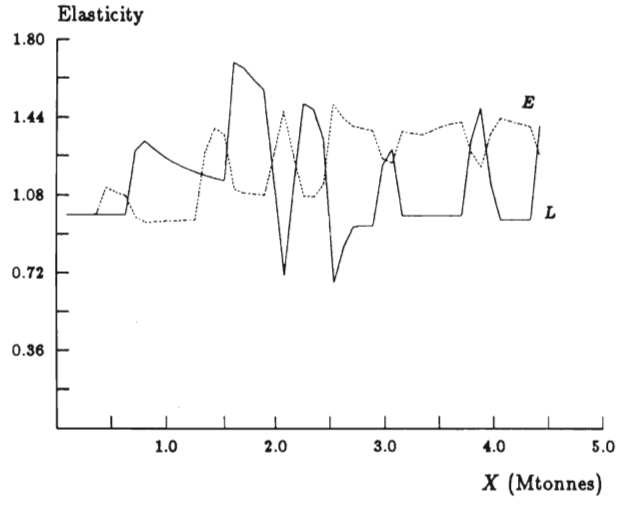


Figure 8.7: Demand elasticities of labour and energy in 1974. (Actual 1974 prices.)

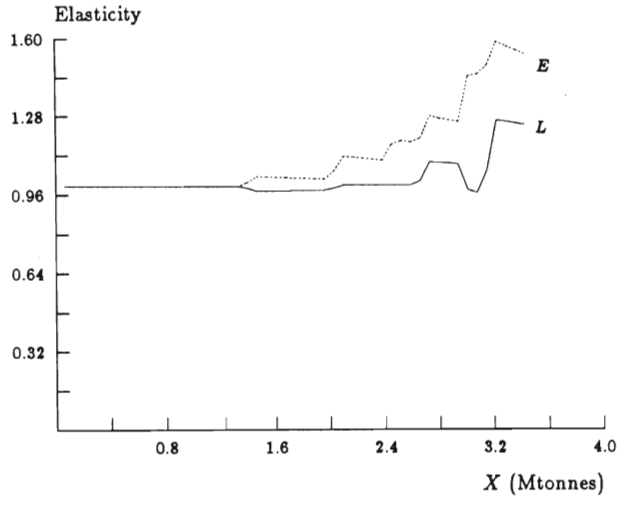


Figure 8.8: Demand elasticities of labour and energy in 1979. (Actual 1979 prices.)

Productivity change

For all years the distance between the isoquants in Figure 8.4 is 500 ktonnes and the scale on the axis is the same during the entire period. The productivity improvements can be seen by following any isoquant representing the same output level from year to year. In Figure 8.4 three isoquant levels are indicated by arrows, 500, 1500 and 2500 ktonnes, respectively.

For all levels there is a marked movement towards the energy axis. There is also a substantial shift towards the origin, which is somewhat stronger the higher the levels of output. The long-run effect of ex ante substitution possibilities through exploitation of economies of scale, particularly between capital and labour, and energy saving by the introduction of new dry processes has resulted in west-southwest movements of the isoquants.

Another informative visualisation when studying the change of the short-run production function is to look at the development of the transformed isoquant map of the short-run function into the input-coefficient space. A transformation of the isoquant maps in Figure 8.4 (except for 1960) is shown in Figure 8.9.

The transformed isoquant map of the short-run function, called the capacity region, shows the region of feasible input coefficients of the industry production function as a whole. Thus, this region must necessarily be narrower than the capacity distribution region portraying the individual units. The boundary towards the origin of the feasible region is called the efficiency frontier.[6]

The west-southwest movement of the feasible region is clearly exhibited. For 1979 the region almost collapses into two lines. In general the right hand outgrowth represents the least efficient kiln; for 1979 the right hand branch represents the remaining wet capacity.

Substitution properties

Figure 8.4 reveals a general tendency for the isoquants to become steeper over the years, i.e., the scope for labour substitution diminishes relative to the scope for energy substitution. However, since the isoquants consist of piecewise linear segments it is difficult to find numerical measures confirming this visual impression.

The conventional measure of substitution, the elasticity of substitution, is zero at the corner points and infinite along the segments. One possi-

[6] See Chapters 3 and 5.

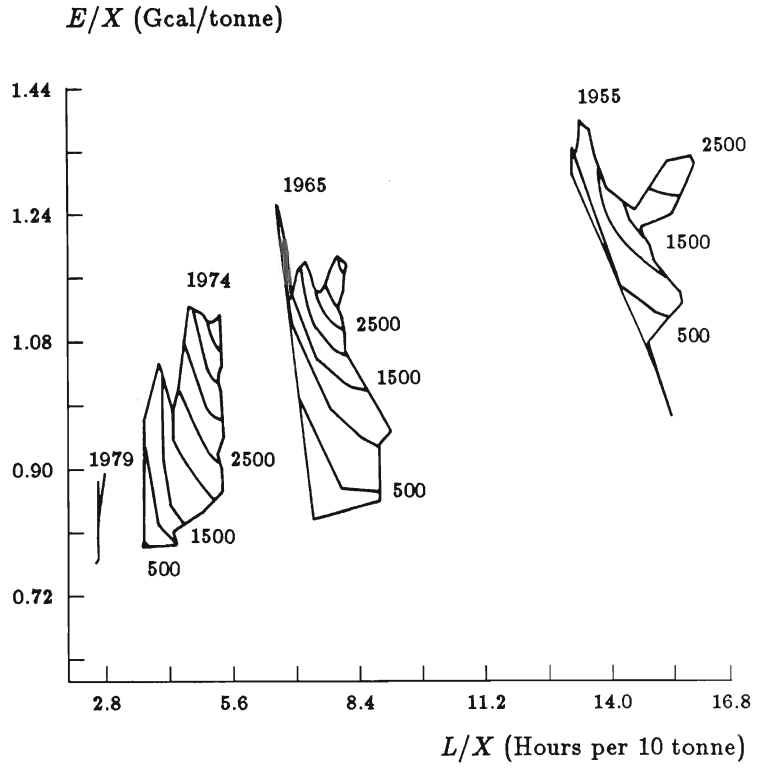


Figure 8.9: The development of the capacity region of the short-run industry production function.

bility is to compute an arc elasticity directly by calculating the ratio between the percentage change in the factor ratio and the percentage change in the slope for two consecutive isoquant segments as defined in Equation (5.18).

The arc elasticities of substitution for an output level of 1500 ktonnes for all years are listed in Table 8.3. The number of isoquant segments varies from year to year, and the number of arc elasticities is equal to this number less 1. Although there are many very low values in Table 8.3, the values vary considerably, up to quite high values, and it is difficult to read off any systematic pattern. Thus, Hildenbrand's [1981] claim that as a

Table 8.3: Arc elasticities of substitution. (Output level 1500 ktonnes.)

No.*	1955	1960	1965	1970	1974	1979
1	**	**	0.10	0.03	2.88	0.03
2	0.02	**	16.23	4.15	0.05	
3	0.01	0.03	0.24	0.98	0.38	
4		0.02	0.06	0.43	0.03	
5		4.72	0.35	0.02	0.16	
6		0.05	7.23	0.06		
7		0.03	0.04	0.26		
8		0.04	0.76	123.74		
9		0.02	1.06			
10		0.13	0.03			
11			1.62			

* isoquant segment pair number from upper boundary
** virtually vertical isoquant segment

"general empirical fact" (his quotation marks) this elasticity is quite low is not confirmed here.[7]

Technical advance and bias measures

Figures 8.4–8.6 and 8.9 give a picture of significant change for the short-run production function. With respect to numerical measures of the changes, we shall here adopt Salter's measures of technical advance and factor bias,[8] set out in Tables 8.4 and 8.5.

We have chosen to utilise 1979-prices (the Paasche index), and have calculated the degree of technical progress and the factor bias for the three output levels marked out in Figures 8.4 and 8.9, 500, 1500 and 2500 ktonnes, in addition to 3500 ktonnes and the frontier of the capacity region shown in Figure 8.9. The short-run industry function program provides us with the current unit costs C, along the expansion path, corresponding to the 1979 prices.

[7] See also Appendix 5.1.

[8] See Section 3.6.

Table 8.4: The Salter technical advance measure T in 1979 prices.

$$T = \frac{C_{t+1}}{C_t}\Big|_{X=X_0} \text{ and } C_t = \text{minimised unit cost in year } t.$$

Year	Frontier	Output levels, X, in ktonnes			
		500	1500	2500	3500
1955/60	0.84	0.82	0.83	0.82	
1960/65	0.74	0.79	0.80	0.78	
1965/70	0.82	0.78	0.78	0.82	0.83
1970/74	0.90	0.89	0.91	0.93	0.94
1974/79	0.90	0.89	0.82	0.74	0.76
1955/79	0.41	0.40	0.38	0.36	

Table 8.5: The Salter factor bias measure D_{EL} in 1979 prices.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}}\frac{L_{t_1}}{L_{t_2}}\Big|_{X=X_0} \text{ and } t_1 < t_2.$$

Year	Frontier	Output levels, X, in ktonnes			
		500	1500	2500	3500
1955/60	2.01	1.58	1.13	1.15	
1960/65	0.88	0.85	1.43	1.51	
1965/70	1.82	1.65	1.33	1.38	1.32
1970/74	1.04	1.20	1.02	1.90	1.03
1974/79	1.31	1.30	1.46	1.55	1.58
1955/79	4.41	3.41	3.19	3.36	

The current unit cost reduction from 1955 to 1979 was around 60 per cent, and increased from 59 to 64 per cent while moving from the frontier (i.e. the boundary towards the origin and the axes in Figure 8.9) to higher output levels. This way of measuring technical advance confirms and quantifies the impressions from Figure 8.4 that technical progress has been rapid between 1960 and 1965, particularly on the frontier with a unit cost reduction of 26 per cent due to the introduction of new kilns. Between 1970 and 1974, and 1974 and 1979 the technical advance slowed down markedly on the frontier, and during these periods, technical advance stemmed from increases in labour productivity. The advance measures for 1979/74 show the gain for the industry from the rest of the kilns catching up with the best practice techniques. There are substantial cost reductions for higher total output levels.

Generally the factor bias measures show a strong labour-saving bias, except at the frontier 1965/60 due to the northwest-southeast extension of the frontier and the 500 ktonnes isoquant in 1960, and on the 2500 ktonnes isoquant in 1974/70 due to the changed slope of the isoquants.) The optimal energy-labour ratio has increased three to four times between 1955 and 1979. The results vary somewhat between different pairs of years and for different isoquant levels. The change between 1970 and 1974 has been the smallest.

Both the advance and the bias measures depend on the prices chosen. In order to check the sensitivity of the results the measures have also been calculated for 1955 prices (by using the Laspeyre index). The same pattern for technical advance results but on a somewhat lower level (e.g., cost reduction 1979/55 0.50–0.42), which is to be expected since relatively speaking the price of labour and labour productivity have increased the most between 1955 and 1979. However, the overall picture for the bias measure is the same as for 1979 prices.

8.5 Technology

As pointed out in Section 8.2 cement manufacturing has undergone a transformation from an all-wet technology in 1955 to an almost all-dry technology in 1979. The first step in this direction was taken by the introduction of the semi-dry (Lepol) technology in 1960. The utilisation pattern of these new technologies is shown for the years 1960, 1970 and 1979 in Figures 8.10 and 8.11, by means of the partial utilisation strips introduced in Section 5.3.

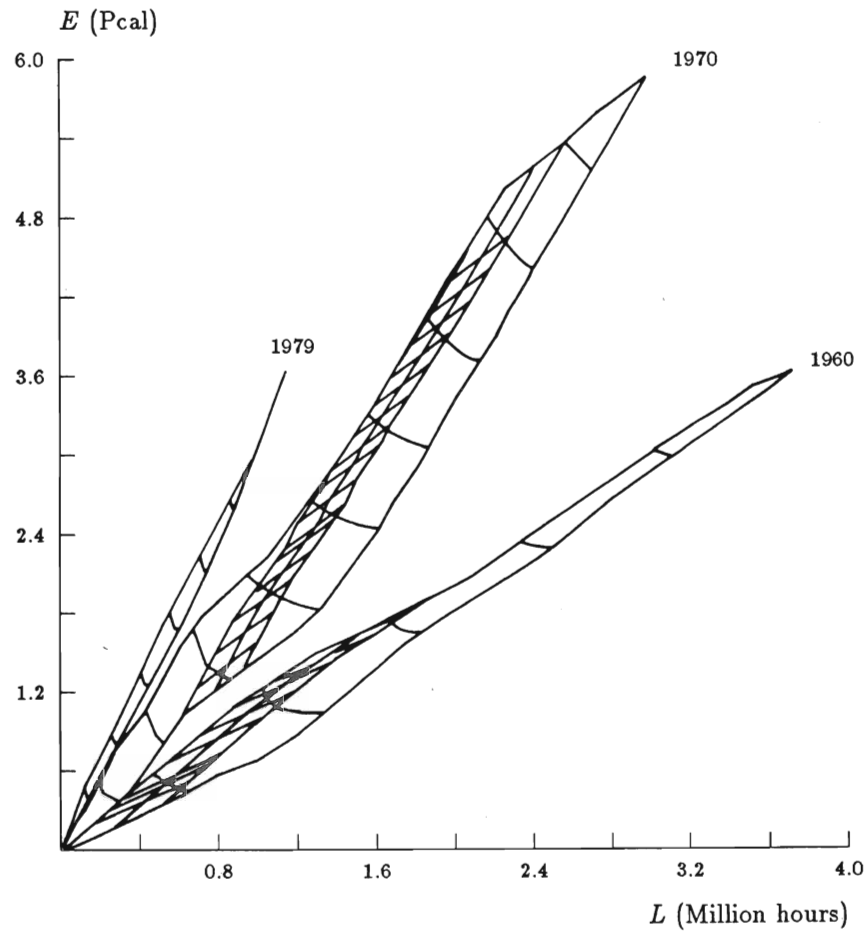


Figure 8.10: Utilisation pattern of semi-dry kilns in 1960, 1970 and 1979.

In 1960 the two semi-dry kilns were the most efficient with respect to energy. As seen in Figure 8.10 the utilisation pattern is to some extent dependent on the relative price in 1960, and to a much larger extent in 1970. In both years they would have been fully utilised at low levels of

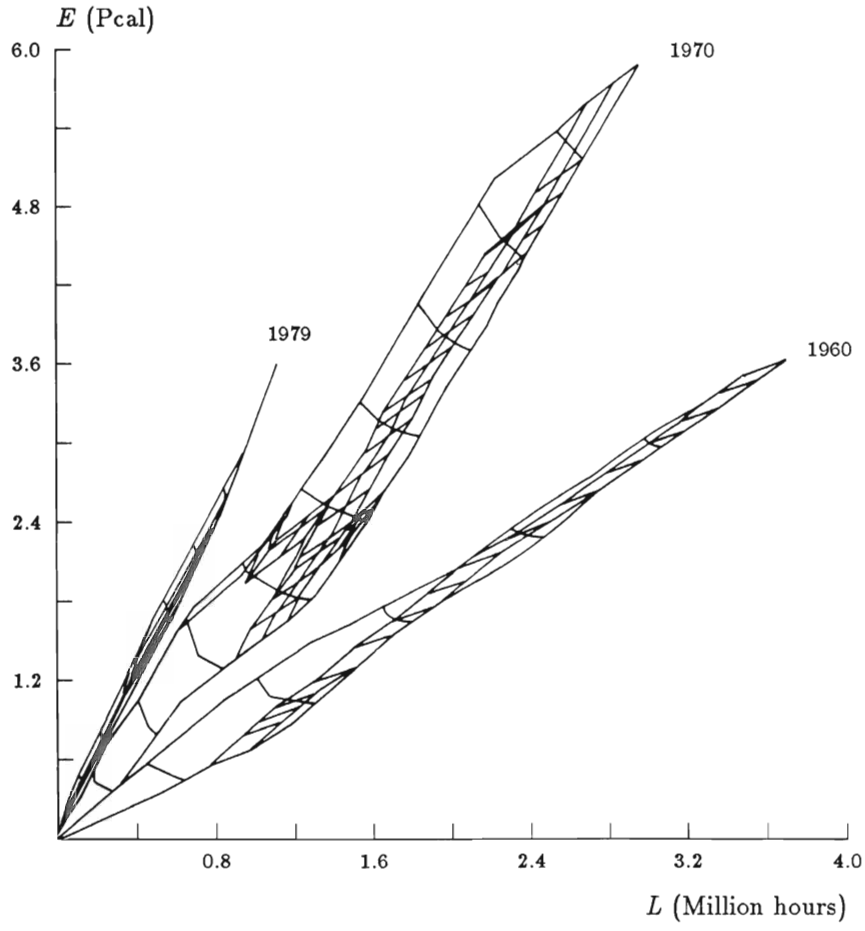


Figure 8.11: Utilisation pattern of dry kilns in 1960, 1970 and 1979.

capacity utilisation when the relative price of energy was sufficiently high. In 1960 the semi-dry kilns were used right from the beginning while in 1970 some wet kilns were the more efficient.

The one dry kiln appearing in 1960 was not the most efficient in the use

of energy, but was very close to best-practice. With respect to labour, however, it had the highest input coefficient. These features result in the utilisation pattern being very relative-price-dependent, as seen in Figure 8.11. This kiln is not fully utilised independently of relative price until all capacity of the entire industry is exhausted. In 1970 a twin kiln has appeared showing the same utilisation pattern as in 1960. Two more small dry kilns appear in 1970 showing a quite different utilisation pattern, the utilisation being very scale-dependent. It is interesting to note that it is not until the late 1970s that the potential energy efficiency of the dry technology was realised, as is indicated by the shifts of the isoquants. The two semi-dry kilns were closed down and only three wet ones remained.

8.6 Structural features

The short-run cost function

The Salter technical advance measure utilises just a few points on the current average cost curves. The complete average and marginal cost curves provide us with a comprehensive picture of the change of the variable cost structure over time. The average and marginal cost curves are shown in Figure 8.12.

The difference in absolute cost levels reflects the values of the Salter technical advance measure in Table 8.4. The average cost curves increase very slowly and smoothly in all years and are almost flat in 1979. The figure clearly shows that the Salter measure is fairly independent of the output levels chosen.

The marginal cost curves provide us with a more detailed and richer structural description. In 1955 there is a marked J-shaped tail to the marginal cost curve, reflecting the upward pointing protuberance of the capacity region in 1955 as shown in Figure 8.9. In 1974 the marginal cost curve is characterised by a marked step after 30 per cent of the capacity has been exhausted. After this level, the marginal cost curve develops almost parallel to the average cost curve without any upward turning tail at the end. The first flat portion of the curve reflects the location of the three most efficient plants shown in the capacity distribution.[9] In 1979 the two best-practice plants constituted about 60 per cent of the capacity reflected

[9] See Figure 8.3.

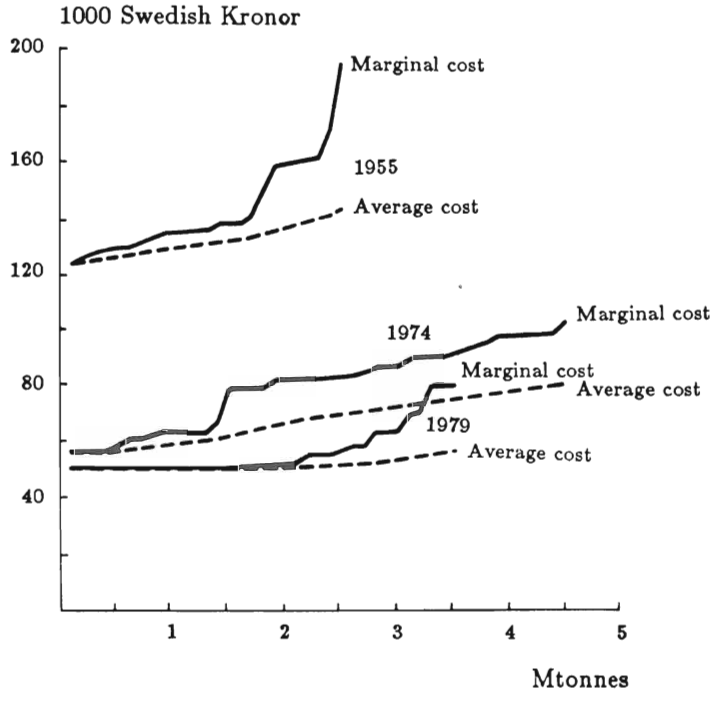


Figure 8.12: The marginal and average cost functions, along the expansion paths, for 1955, 1974 and 1979 in 1979-prices.

in the flat portion of the marginal cost curve where it is almost identical to the average cost curve. The upward pointing tail of the marginal cost curve for the last 40 per cent of the capacity corresponds to the distribution of energy input coefficients shown in Figure 8.2.

Elasticity of scale

The evenness of the structure can also be illustrated by the spacing of the isoquants, measured, for instance, by the development of the elasticity of scale along a factor ray. Note that the elasticity of cost, calculated as the

ratio between the marginal and average costs and shown in Figure 8.12,[10] can no longer be interpreted as the inverse of the elasticity of scale, since elasticity of scale does not exist uniquely at the isoquant corners and the isocline consists only of corner points. We must therefore choose another basis for calculating the scale elasticity.[11]

In Table 8.6 the development of the elasticity of scale, ε, is shown for the average factor ratio for each isoquant level. When the factor ray is outside the substitution region we have chosen the values of the scale elasticity of the bordering isoquant segment in question.

As discussed in Section 5.4, ε does not necessarily decrease monotonically with increasing output. Even though the general tendency is for values to decrease, we observe also increasing phases of the scale elasticity for all years except 1979.

Even if the elasticity of scale is calculated along a factor ray, it turns out that the values shift downwards at the same output levels at which the corresponding marginal cost curves shift upwards in Figure 8.12. The impact of the best-practice units in 1979 for the industry performance is clearly exhibited by the almost unity values of the scale elasticity corresponding to the flat part of the average cost curve in Figure 8.12.

For the variation of the scale elasticity along isoquants our results indicate that it is rather limited. Thus the general tendency of the results in Table 8.6 is fairly independent of the chosen factor ray.

In Hildenbrand [1981] there is a general statement that the short-run function cannot be homothetic. However, in our case some of the years with narrow substitution regions may be considered as approximations. Two tests of homotheticity are the shape of isoclines and the values of the scale elasticity along an isoquant. Figure 5.2 shows one isocline for 1974 corresponding to the average of the observed prices. Although the isocline is not a ray through the origin, linearity might be a good approximation over some sections of the substitution region. In Table 8.7 scale elasticity values for the 1500 ktonnes isoquant for all years are shown. These values do not vary that much along this isoquant.

Efficiency

In analogy to the structural efficiency measures introduced in Section 3.4, structural efficiency measures for the short-run function may be obtained

[10] As in Hildenbrand [1981].

[11] See Section 5.4.

Table 8.6: The development of the scale elasticity along the average factor rays.

Year	Output levels in ktonnes									Average factor ratio
	500	1000	1500	2000	2500	3000	3500	4000	4500	
1955	0.99	0.95	0.96	0.84	0.85					0.087
1960	1.00	0.96	0.97	0.90	0.88					0.098
1965	0.99	0.96	0.97	0.92	0.92	0.84	0.83			0.147
1970	0.96	0.92	0.93	0.84	0.86	0.83	0.83	0.83	0.83	0.197
1974	0.93	0.88	0.91	0.81	0.83	0.84	0.84	0.83	0.79	0.210
1979	1.00	0.99	0.99	0.99	0.99	0.92	0.81			0.315

Table 8.7: The development of the scale elasticity along an isoquant. (Output level 1500 ktonnes.)

Isoquant segment no.	Years					
	1955	1960	1965	1970	1974	1979
1	*	*	0.92	0.78	0.83	0.99
2	0.94	*	0.96	0.92	0.84	1.00
3	0.84	0.91	0.96	0.92	0.91	
4	0.84	0.91	0.97	0.93	0.79	
5		0.93	0.92	0.87	0.78	
6		0.93	0.93	0.80		
7		0.93	0.93	0.82		
8		0.91	0.91	0.82		
9		0.91	0.91			
10		0.88	0.91			
11		0.90	0.88			
12			0.88			

* Virtually vertical isoquant segment.

by comparing observed total inputs to potential inputs with the same factor ratio on the short-run function at the observed output, or by comparing observed output to potential output on the short-run function employing the observed amount of inputs. The former approach is followed here, and moreover the degree of adjustment of input proportions to relative prices is also measured. We must again remember that important factors in the real industry optimisation are excluded here, particularly transport costs.

By comparing "actual" costs (i.e., costs imputed by the observed average input prices for the respective years) with the costs of producing the same output with the same observed factor ratio on the short-run function, a measure analogous to Farrell's measure of technical efficiency is obtained. By further comparing these last costs with the minimal cost along the isoquant corresponding to actual observed output we obtain a measure analogous to Farrell's measure of price, or allocative efficiency. The product of these measures yields Farrell's overall efficiency measure.

Table 8.8: Estimates of efficiency.

Year	Technical efficiency	Allocative efficiency	Overall efficiency
1960*	0.98	1.00	0.98
1970	0.95	0.998	0.95
1974	0.97	0.99	0.97
1979	0.88**	0.97***	

* 1955 output is equal to capacity and the efficiency measures are equal to one and in 1965 observed output exceeds capacity.

** Since the observed average factor ratio lies outside and above the isoquant for the observed output level, we have compared observed costs with the computed costs at the boundary corresponding to the observed output level. Thus this measure is not a true Farrell measure of technical efficiency.

*** The minimum costs are compared with the costs at the border of the same isoquant.

The values of the efficiency measures are shown in Table 8.8.[12] For all years except 1979 the efficiency values are very high particularly taking into consideration that transport costs are excluded.

This is most surprising, particularly with regard to 1974 where the low degree of capacity utilisation should have affected the technical efficiency value downwards. The adjustment to relative prices is almost perfect even in 1974 with its considerably wider region of substitution. The overall efficiency measure indicates that a "perfect" optimisation should have yielded less than 3 per cent cost reduction in 1974 in spite of a very low capacity utilisation. One reason for this high efficiency level is that the Swedish cement industry in 1974 was a monopoly with an elaborate production model for short-run optimisation. In 1979 the relatively low value of technical efficiency was due to the very low degree of capacity utilisation of the largest unit which came on stream that year. Since capacity utilisation was about 100 per cent or more in 1955 and 1965, it is not possible to calculate the defined structural efficiency measures for these years.

[12] See Chapter 3.

8.7 Conclusions

In this chapter we have performed an analysis of industrial structure and structural change for an industry consisting of well defined production units.

The empirical results show that the process of structural change of the Swedish cement industry has been characterised by a substitution process from labour towards energy in combination with a rather rapid cost-reducing technical progress. Factors explaining this development are long-run ex ante substitution possibilities and increasing returns to scale between capital and labour-energy when introducing new techniques, and disembodied improvements particularly with respect to labour saving.

With respect to unit cost the flattening out of the cost curve over the period results in a structure very similar to one that appears in the long-run steady state of almost equal production units. From an industrial policy point of view it should be observed that such a structure is very vulnerable. With only a small cost or price change the entire industry may find itself operating at a loss.

9

The Swedish Pulp Industry

9.1 Introduction

The pulp industry has been one of the main industries in Sweden during the last century. The industry is a very large energy consumer, and due to its geographical dispersion the impact it has on regional employment is particularly important. From its very beginning about a hundred years ago, this industry has undergone a gradual, continuous structural transformation. One aspect of this structural change has been the development of different technologies. Another related aspect has been the development of the size distribution of plants. For industrial policy the concept structural rationalisation, discussed in Chapter 2, has been particularly important, since the industry is characterised by typical vintages of capital equipment and embodied technical change. The Swedish pulp industry represents three stages of technological development: the groundwood mills mostly founded in the latter part of the nineteenth century, the sulphite mills from the first two decades of this century and the sulphate mills from the second and the third decades of this century.

In order to analyse the long-run technical change in the twentieth century the short-run industry function will be utilised for each technology for a selected number of years (1920, 1929, 1937, 1954 and 1974). Except for the first two years, the selected data are about 20 years apart. In some analyses the period 1929 to 1954 has been divided into several subperiods.

9.2 Data

The pulp production processes

Mechanical pulp is produced by grinding wood while adding water in order to free the cellulose fibres. Other components of the wood, mainly lignin and hemicelluloses, remain in the pulp. Thus the pulp yield is very high, nearly 100 percent of the dry weight of the wood. The technology was originally borrowed from flour milling.

The motive power of wood pulp mills is electricity and in the decades between the wars some plants were still being driven directly by water power. Even today to some degree the locations of the plants are due to their proximity to power plants. In most respects the milling machinery is rather simple and does not require as much capital as the production of chemical pulp does. A modern version of this process is thermo mechanical pulp.

In the sulphate and sulphite mills the wood is chipped into small pieces, which are boiled under pressure in an alkaline or acid solution. The lower yield compared to wood pulp is due to the fact that the chemical processes dissolve the lignin and hemicelluloses, leaving only the pure cellulose fibres, which represent about half the dry weight of the wood. Since chemical pulps are more or less pure cellulose fibres, they are stronger than wood pulp and can be used for a variety of purposes.

Pine fibres are longer and stronger than spruce fibres. For this reason sulphate pulp is used to produce wrapping paper, board and other products which have been in high demand since the 1950s.

The processes have been highly mechanised during the whole period. Transportation, both to the mills and within them, has become more and more mechanised.

The consumption of energy at chemical pulp mills is large. However, over time it has decreased per unit of output, since it has been possible to reuse the chemicals and the heat. The burning of the waste liquid gives so much heat energy that modern sulphate mills are self-sufficient in this energy. In Table 9.1 the main pulp processes are summarised.

The data set

In this study we have used primary data for individual unintegrated (i.e., not integrated with a paper factory) pulp plants in Sweden. The reason

Table 9.1: Main pulp processes.

Product	Raw material: (solid m^3/tonne)		Technical processes	Examples of final use
1 Mechanical pulp	Spruce	2.6	Grinding	Newsprint
2 Sulphite pulp	Spruce	5.0	Acid cooking	Printing and writing paper
3 Sulphate pulp	Pine	5.0	Alkaline cooking	Wrapping paper
	Birch	4.0		Liner board

for choosing the specific years mentioned above are that they are typical "boom" years for the pulp industry, with almost full capacity utilisation. The data are based on the annual Industrial Statistics at plant level, collected by Statistics Sweden. For 1974, however, the data were collected directly from the individual firms.

The plants are divided into three categories: mechanical pulp, sulphite pulp and sulphate pulp. With a few exceptions, particularly in 1974, the data cover all unintegrated plants for the respective years. The number of units in some years are gathered in Table 9.2.

Table 9.2: Number of pulp plants.

	1920	1937	1954	1974
Mechanical pulp	40	41	15	7
Sulphite pulp	35	34	19	12
Sulphate pulp	12	19	15	12

These plants represent about 50 per cent of total capacity (unintegrated plus integrated plant capacity) for sulphite and sulphate pulp and about 30–40 per cent of total capacity of mechanical pulp, taken as the average for

the period. (These shares decreased markedly in the last sample year.) It is a reasonable approximation to consider the three different pulp categories as homogeneous products. Output is measured in tonnes of pulp produced during the year. Capacity, reflecting potential output of the plants, is also measured in tonnes. The labour input variable is defined as the hours worked by production and maintenance workers. There are two basic types of energy: fuel and electricity. Energy consumption is aggregated to kWhs using energy content.

The following notation will be employed:

L = labour (hours)
E = energy (kWhs)
X = output (tonnes)
L/X, E/X = input coefficients

9.3 Structural description

Due to the large change in input coefficients and size distribution we will limit the presentation of the structural development to Salter diagrams for labour and energy as set out in Figures 9.1–9.6.

The development of the labour-input coefficient distributions is for all three processes remarkably similar. While there was a significant shift during the periods 1920–37 and 1954–74, the 20 year period covering the second world war saw development at a stand still. The J-shaped distributions of the earlier years have gradually become flatter and give an almost even distribution in 1974.

The development of the energy coefficient distributions is quite different for the three technologies. The direct use of water power in the mechanical pulp industry is not registered as energy consumption. This explains the fact that more than 40 per cent of capacity in 1920 did not consume any energy, and it is not until 1954 that this energy source is unimportant. Between 1920 and 1954 the use of electricity has gradually taken over but without any basic changes in the technology. During the period 1954–74 there was a markedly upward shift in the distribution and the 10–20 per cent tail of the distributions has disappeared. This change must be attributed to intrusive mechanisation, and provides a unique example of the substitution process dominating a general productivity increase.

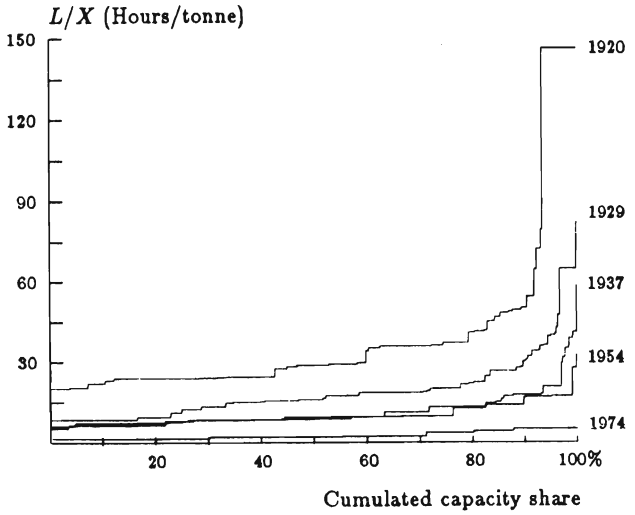


Figure 9.1: The development of the labour-input coefficient distribution for selected years. Mechanical pulp.

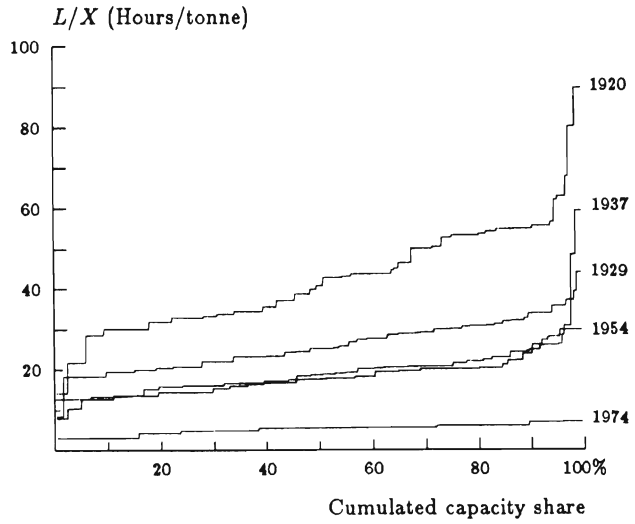


Figure 9.2: The development of the labour-input coefficient distribution for selected years. Sulphite pulp.

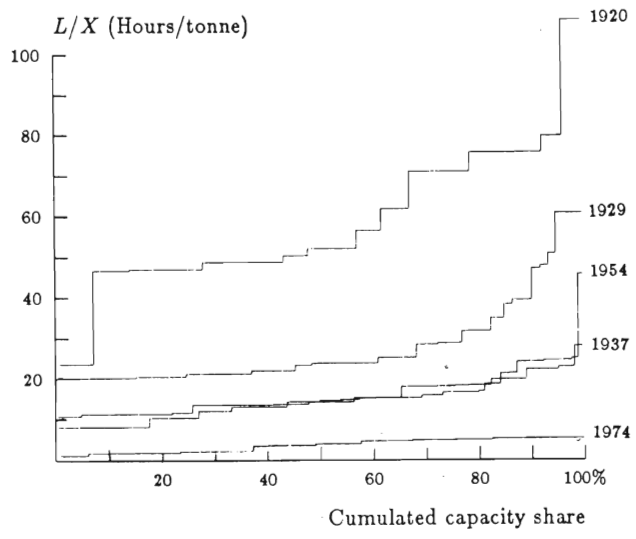


Figure 9.3: The development of the labour-input coefficient distribution for selected years. Sulphate pulp.

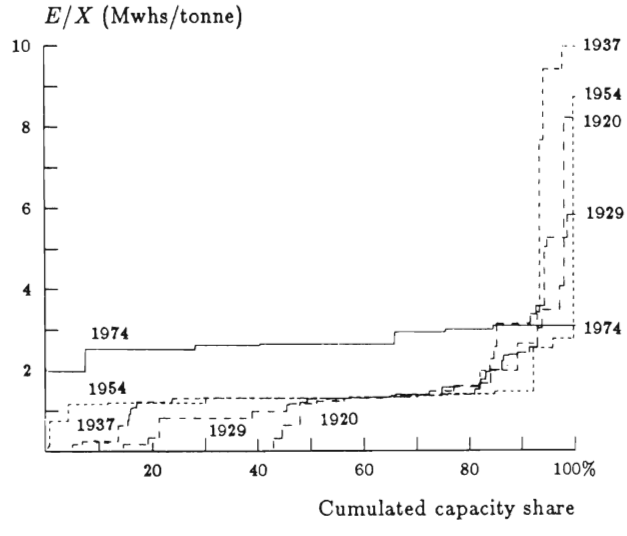


Figure 9.4: The development of the energy-input coefficient distribution for selected years. Mechanical pulp.

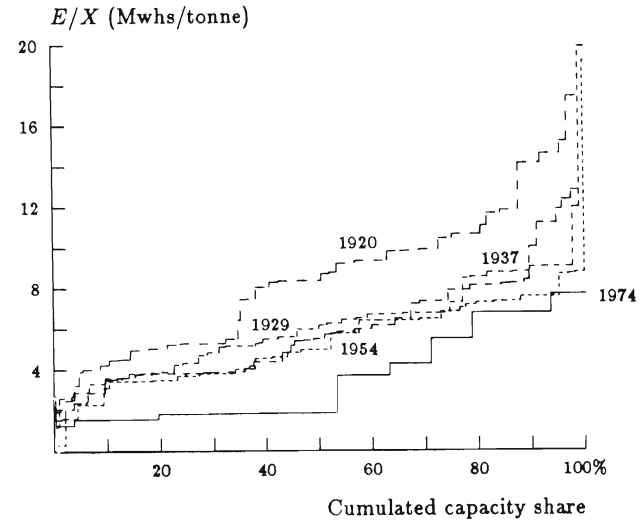


Figure 9.5: The development of the energy-input coefficient distribution for selected years. Sulphite pulp.

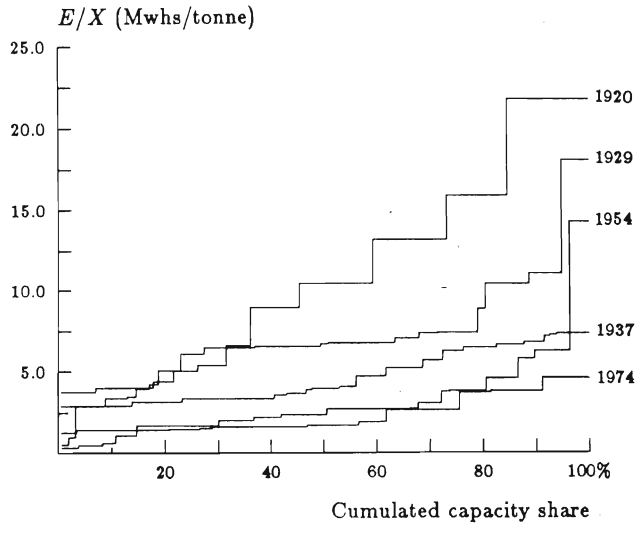


Figure 9.6: The development of the energy-input coefficient distribution for selected years. Sulphate pulp.

The energy-input coefficient distributions for the chemical processes have gradually shifted downwards. The stagnation of output growth for sulphite and a stationary technology resulted in almost constant distributions during the period 1929–54. This does not hold for the expanding sulphate industry. For both technologies the movement of the best practice has been quite limited compared with the movement of the average, and thus the latter is more responsible for the flattening of the distributions.

9.4 The short-run industry production function and technical change

Mechanical pulp

The development of the short-run function shown in Figure 9.7 is characterised by a gradual change of the substitution region from the labour towards the energy axis. The tendency towards more flatter input-coefficient distributions in Figure 9.1 is confirmed here by a more narrow substitution region in 1954 and 1974 than in the earlier years. This is also clearly revealed by the development of the capacity region in Figure 9.8.

In Figure 9.7 the development of one isoquant (that for 150 ktonnes) is marked by arrows. The shift of the isoquant between 1929 and 1974 reveals that technical change has been limited to a substitution between labour and electricity without any overall improvement in productivity. The development of the capacity region shown in Figure 9.8 further underlines this picture. In our experience this is a rather rare case, since technical change in most industries seems to be characterised by a simultaneous process of substitution and overall productivity improvements shifting (although not along a straight line) the isoquants towards the origin.

One explanation for these differences in structural change may be due to the fact that basically, except for size, which has increased, the production process for mechanical pulp has remained unaltered through time. This increase in size has mainly affected the unit requirements of labour, not electricity. On the other hand, for sulphite and sulphate pulp it is easy to find major process innovations which have saved both labour and energy.

We also note, for example, by traversing from the upper starting point to the lower end point of the isoquant for 150 ktonnes of mechanical pulp in 1937, that electricity consumption can be decreased by about 47 per cent if the labour input is increased by 20 per cent. The isoquant is L-shaped and particularly steep at the beginning. As can be calculated, the first four

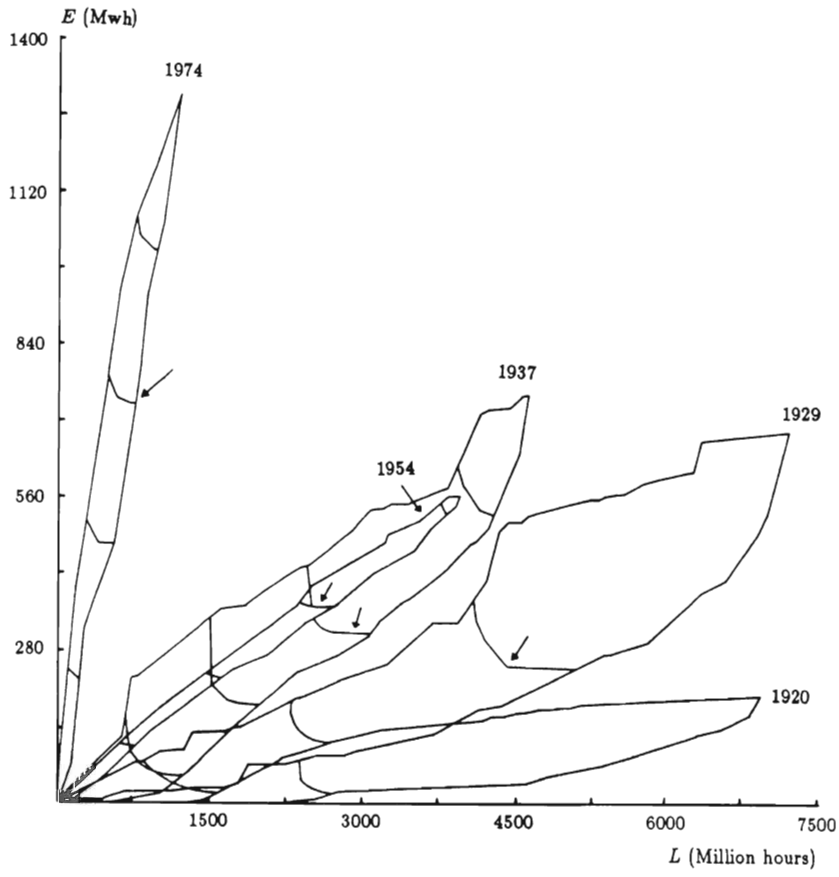


Figure 9.7: The development of the short-run industry production function for mechanical pulp.

line segments of the isoquant reduces electricity input by as much as 44 per cent while at the same time labour input is increased by only 1.7 per cent. On the other hand, for the last 13 segments of the isoquant electricity input can be reduced by only 4.7 per cent when labour input is increased by as much as 18 per cent.

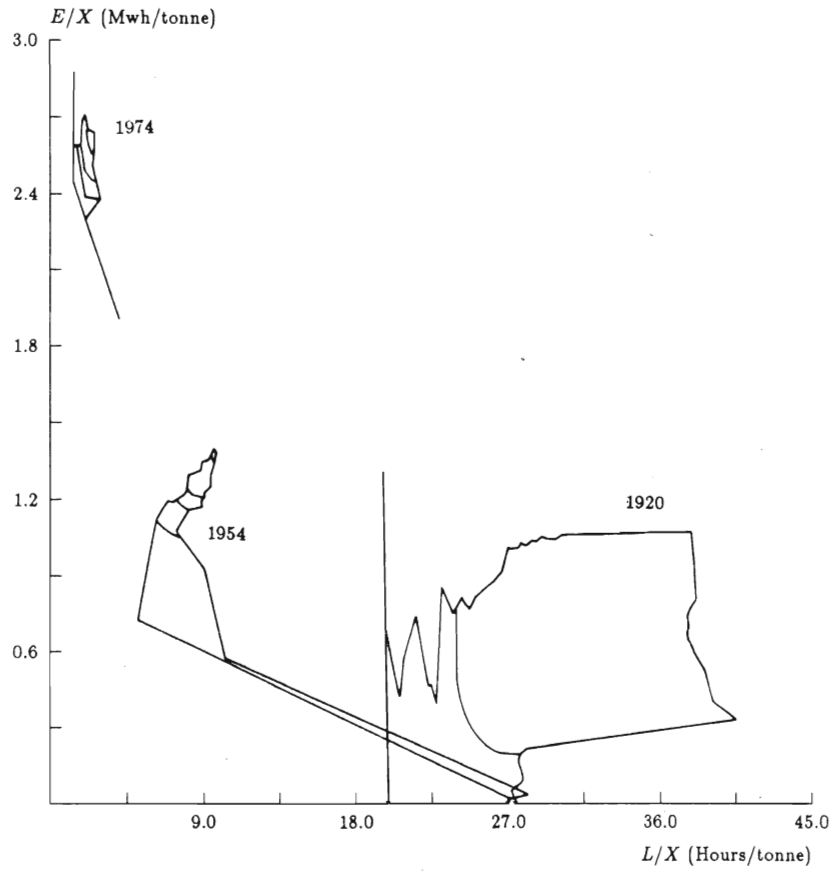


Figure 9.8: The development of the capacity region for mechanical pulp.

Chemical pulp

Figures 9.9–9.12 show the development of the chemical pulp processes and the shifts in one isoquant (500 ktonnes of sulphite pulp and 400 ktonnes of sulphate pulp) are marked by arrows.

The development of both the sulphite and the sulphate processes may be divided into three phases, confirming the impression obtained from the Salter diagrams. During the first phase, 1920 to 1937, both labour and energy productivity increased rapidly, during the second, between 1937

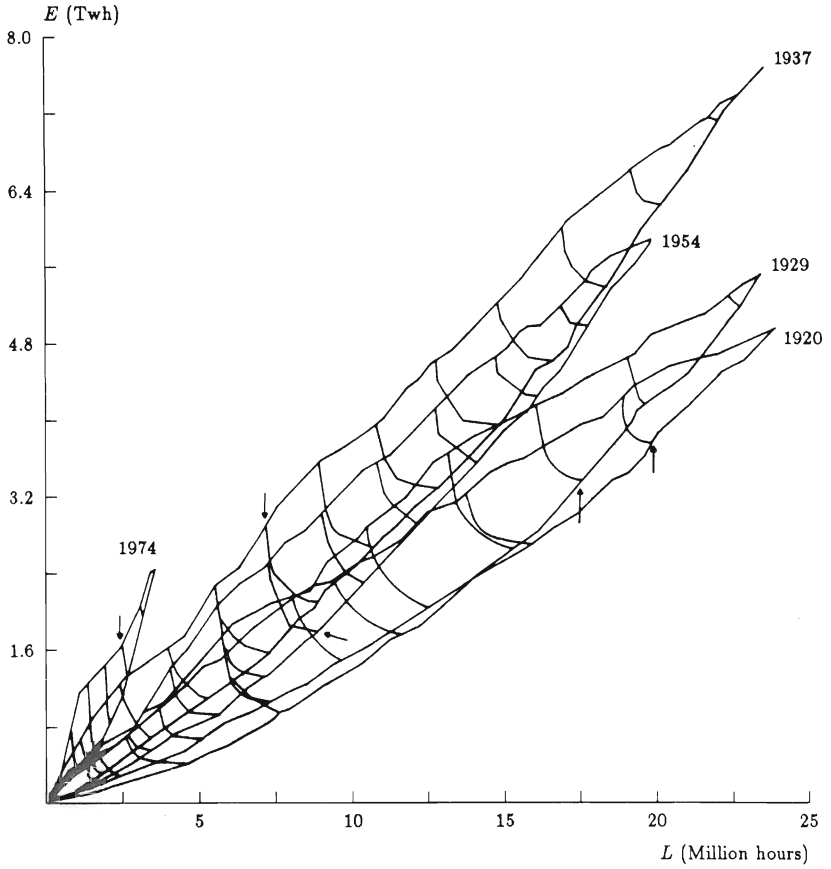


Figure 9.9: The development of the short-run industry production function for sulphite pulp.

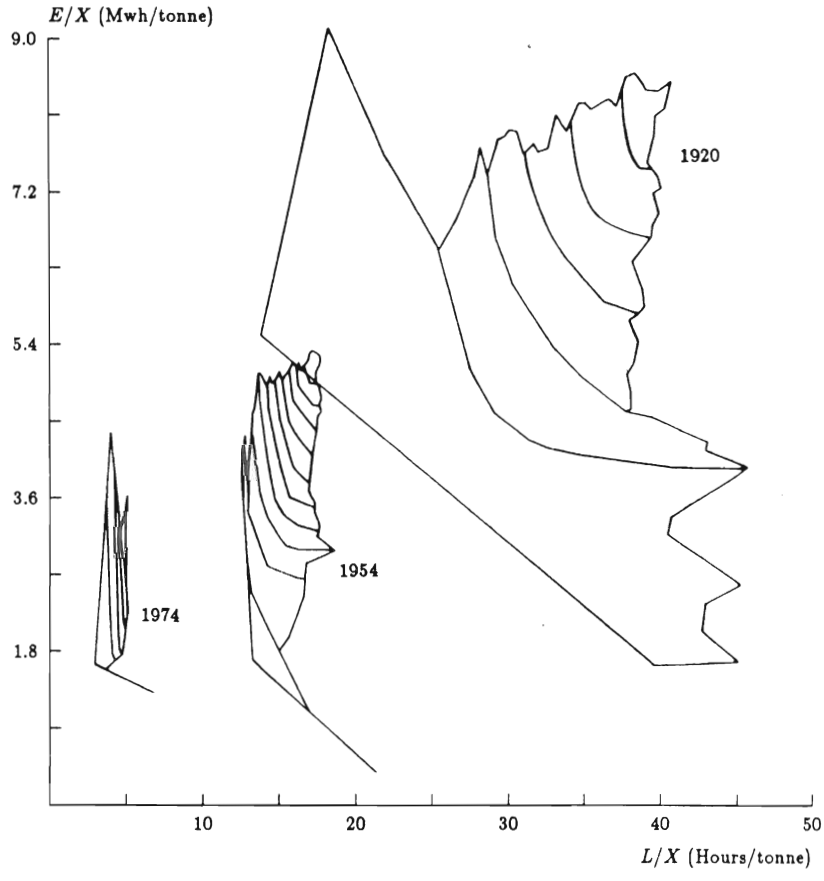


Figure 9.10: The development of the capacity region for sulphite pulp.

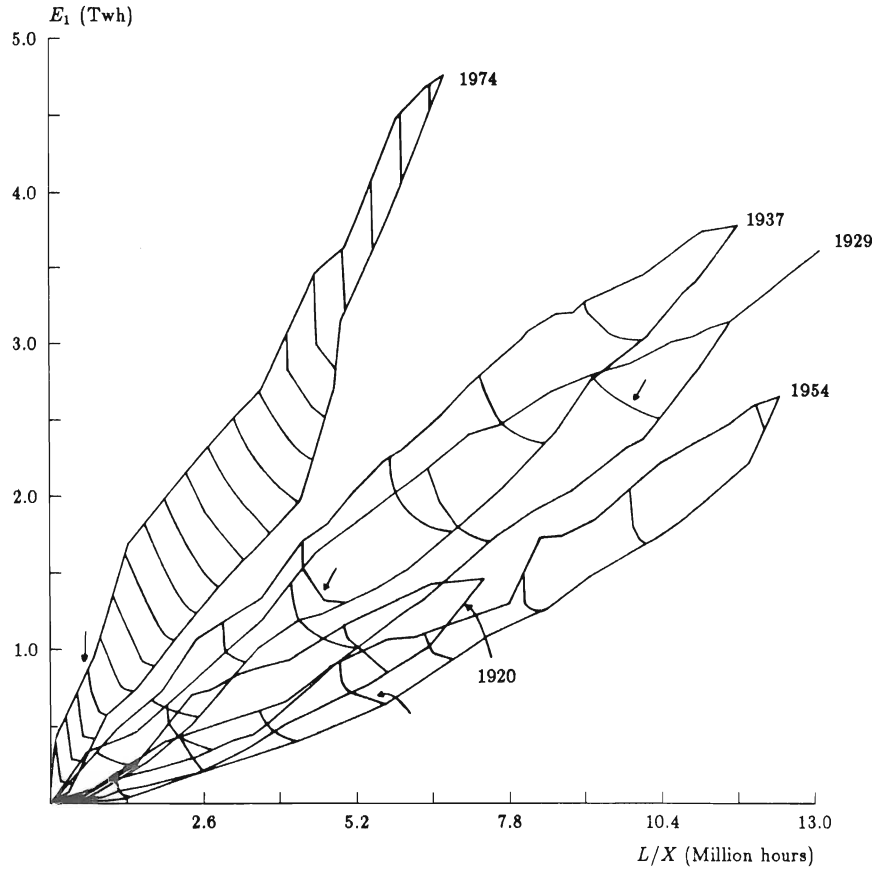


Figure 9.11: The development of the short-run industry production function for sulphate pulp.

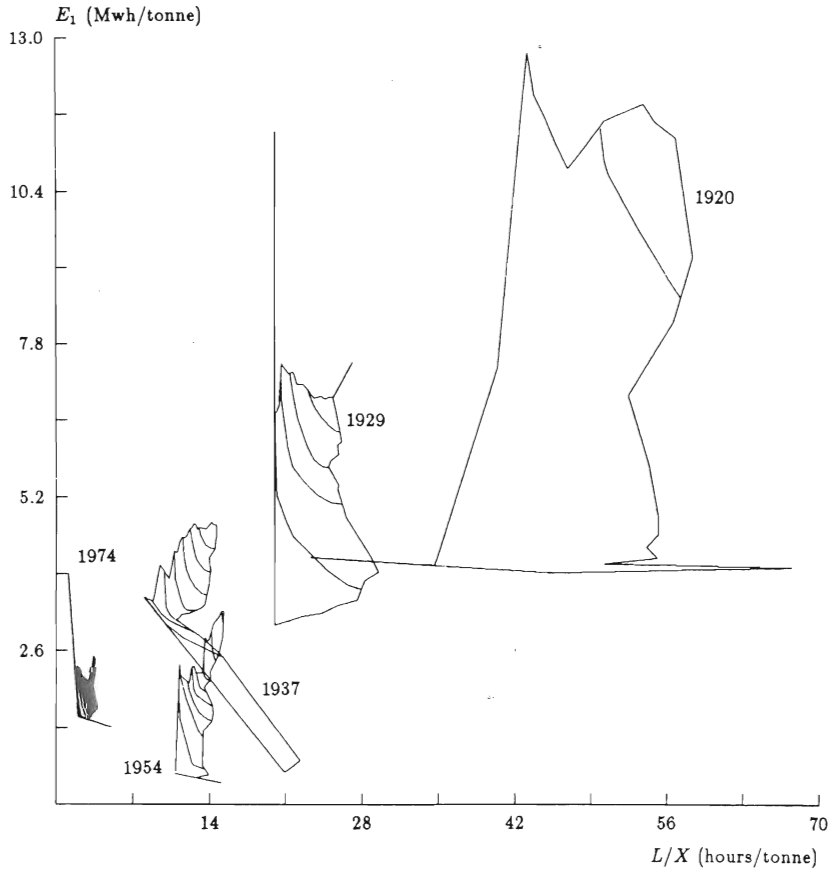


Figure 9.12: The development of the capacity region for sulphate pulp.

and 1954, energy was substituted for labour, while during the last period, 1954 to 1974, an overall productivity improvement took place.

The movement of the marked isoquants in Figures 9.9 and 9.11 follows a lightning-shaped path. The generality of such a movement for all isoquants is clearly brought out in Figure 9.12, which reveals the change in the capacity region. (In Figures 9.8 and 9.10 the intermediate years 1929 and 1937 are not shown in order to enhance the clarity of the exposition.)

The characterisation of technical change

Salter measures of technical change for the sulphite industry are reported in Table 9.3. The pace of progress is somewhat faster in the 1920s than in the 1930s. During the war-years, and particularly the first part, the industry experiences a marked cost increase. Technical progress picks up again in the early 1950s and is the most rapid in the 1960s. For the 50-year period as a whole technical progress has implied cost reductions of the magnitude of 80 to 90 per cent.

As to the nature of the technical change we see from Table 9.4 the overall is one of labour-saving technical change, the exception being the war-time period where we observed cost increases. Labour biased technical change restarts again when technical progress reappears in the early 1950s.

The sulphate industry follows roughly the same pattern as the sulphite industry, as can be seen from Tables 9.5 and 9.6. The decade around the war is characterised by both cost increases and a reversion to labour using technical change. For the last twenty years technical progress is quite strong, resulting in an overall cost reduction for the 50 year period of up to 98 per cent. The overall labour-saving bias is markedly stronger for sulphate pulp than for sulphite. This is to a great extent explained by the investment in large units producing sulphate pulp in the 1960s.

The development of technical change for mechanical pulp is of the same basic pattern as for sulphite and sulphate, but the cost increases during the war years are much smaller, corresponding to a smaller reversal of the factor bias, as seen in Tables 9.7 and 9.8. The overall progress for the 50-year period is of the same magnitude as for sulphate, while the progress in the period 1954–74 is not as great but more equal to sulphite. The labour saving bias of mechanical pulp is, as for sulphate, particularly strong in the same period.

Table 9.3: The Salter technical advance measure T in 1974 prices for sulphite pulp.

$$T = \frac{C_{t+1}}{C_t}\Big|_{X=X^0}, \; C_t = \text{ minimised unit cost in year } t.$$

Year	Frontier	Output levels X in ktonnes											
		100	200	300	400	500	600	700	800	900	1000	1100	1200
1920/29	0.60	0.66	0.63	0.58	0.55								
1929/32	1.86	1.03	0.98	0.97	0.95	0.94	0.91	0.89	0.88	0.87			
1932/33	0.56	0.74	0.83	0.86	0.87	0.88	0.88	0.89	0.90	0.90	0.90		
1933/37	0.90	0.82	0.80	0.82	0.83	0.83	0.84	0.85	0.85	0.84	0.82		
1937/43	2.07	1.70	1.67	1.63	1.66	1.69	1.72	1.72	1.72	1.74	1.79	1.83	1.87
1943/46	0.82	0.80	0.79	0.77	0.75	0.73	0.72	0.72	0.72	0.72	0.71	0.70	
1946/51	1.04	0.99	0.93	0.96	0.96	0.96	0.95	0.93	0.92	0.91	0.94		
1951/52	0.94	0.98	1.01	1.02	1.02	1.02	1.03	1.04	1.05	1.06	1.03		
1952/54	0.95	0.92	0.88	0.84	0.83	0.84	0.84	0.83	0.83	0.83	0.82		
1954/74	0.25	0.24	0.30	0.32	0.34	0.34	0.34						
1920/74	0.22	0.12	0.13	0.14	0.14	0.13							

Table 9.4: The Salter factor bias measure D_{EL} in 1974 prices for sulphite pulp.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}} \frac{L_{t_1}}{L_{t_2}}\Big|_{X=X^0}, \; t_1 < t_2.$$

Year	Frontier	Output levels X in ktonnes											
		100	200	300	400	500	600	700	800	900	1000	1100	1200
1920/29	0.60	2.43	1.06	1.06	1.23	1.13	1.23						
1929/32	1.86	0.55	1.89	1.62	1.33	1.44	1.35	1.27	1.26	1.29	1.27		
1932/33	1.75	1.23	1.01	1.07	1.04	1.01	1.02	1.12	1.16	1.13	1.15		
1933/37	0.98	0.96	1.15	0.95	1.02	1.08	1.12	1.00	0.99	1.04	1.13		
1937/43	0.40	0.45	0.59	0.84	0.85	0.81	0.85	0.93	0.96	0.92	0.86	0.82	0.80
1943/46	1.07	1.08	1.03	0.98	0.86	0.88	0.90	0.84	0.79	0.85	0.91	0.98	
1946/51	0.61	0.79	0.92	1.00	1.09	1.20	1.10	1.18	1.23	1.19	1.17		
1951/52	1.04	0.97	1.00	0.89	1.06	0.96	0.94	0.93	0.90	0.95	0.97		
1952/54	1.24	1.41	0.98	1.17	1.07	1.05	1.05	1.04	1.06	1.02	0.97		
1954/74	1.75	1.65	3.43	2.71	2.06	1.94	1.99						
1920/74	1.35	2.07	3.73	3.82	3.23	2.99							

Table 9.5: The Salter technical advance measure T in 1974 prices for sulphate pulp.

$$T = \frac{C_{t+1}}{C_t}\Big|_{X=X^0}, \quad C_t = \text{minimised unit cost in year } t.$$

Year	Frontier	Output levels X in ktonnes							
		100	200	300	400	500	600	700	800
1920/29	0.86	0.40							
1929/32	0.58	0.61	0.67	0.68	0.67				
1932/33	0.93	0.87	0.84	0.91	0.93	0.92	0.89		
1933/37	0.74	0.74	0.75	0.74	0.74	0.75	0.74		
1937/43	1.52	1.74	1.72	1.67	1.60	1.56	1.60	1.70	
1943/46	0.97	0.91	0.96	0.95	0.95	0.95	0.94	0.88	
1946/51	0.76	0.80	0.88	0.88	0.88	0.88	0.86	0.86	
1951/52	1.27	1.42	1.22	1.16	1.13	1.10	1.09	1.09	
1952/54	0.95	0.78	0.73	0.75	0.76	0.77	0.78	0.77	
1954/74	0.09	0.09	0.12	0.13	0.13	0.13	0.13	0.13	0.13
1920/74	0.04	0.02							

Table 9.6: The Salter factor bias measure D_{EL} in 1974 prices for sulphate pulp.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}} \frac{L_{t_1}}{L_{t_2}}\Big|_{X=X^0}, \; t_1 < t_2.$$

Year	Frontier	Output levels X in ktonnes							
		100	200	300	400	500	600	700	800
1920/29	3.19	1.43							
1929/32	0.79	1.43	1.27	1.23	1.44				
1932/33	0.90	0.85	1.01	0.93	0.87	0.93	0.98		
1933/37	1.07	1.07	0.88	1.03	1.05	1.05	1.07		
1937/43	0.07	0.60	0.76	0.82	0.71	0.66	0.69	0.96	
1943/46	1.23	1.23	0.80	0.63	0.68	0.68	0.80	0.69	
1946/51	14.12	0.80	0.82	0.95	1.14	1.08	0.89	0.87	
1951/52	0.06	0.51	0.70	0.77	0.74	1.03	1.11	1.04	
1952/54	1.49	1.37	1.43	1.14	1.16	0.89	0.90	0.97	
1954/74	83.88	22.25	10.68	9.11	7.18	7.33	6.86	6.18	5.30
1920/74	22.24	17.34							

Table 9.7: The Salter technical advance measure T in 1974 prices for mechanical pulp.

$T = \left.\frac{C_{t+1}}{C_t}\right|_{X=X^0}$, $C_t =$ minimised unit cost in year t.

Year	Frontier	Output levels X in ktonnes							
		50	100	150	200	250	300	350	400
1920/29	0.41	0.36	0.36	0.37					
1929/32	1.15	1.15	1.08	0.97	0.91	0.87	0.85	0.84	
1932/33	0.86	0.88	0.88	0.88	0.87	0.88	0.90	0.93	
1933/37	0.73	0.81	0.83	0.84	0.82	0.80	0.79	0.76	
1937/43	1.00	1.03	1.16	1.18	1.23	1.39	1.58	1.74	
1943/46	0.85	0.73	0.78	0.81	0.82	0.75	0.66	0.60	
1946/51	1.04	1.33	1.14	1.06	1.04	1.05	1.02	1.03	
1951/52	1.30	1.08	1.08	1.11	1.12	1.11	1.13	1.16	
1952/54	0.78	0.83	0.82	0.84	0.83	0.83	0.80	0.78	
1954/74	0.23	0.20	0.19	0.18	0.19	0.20	0.21	0.20	0.20
1920/74	0.06	0.05	0.05	0.04					

Table 9.8: The Salter factor bias measure D_{EL} in 1974 prices for mechanical pulp.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}} \frac{L_{t_1}}{L_{t_2}}\Big|_{X=X^0}, \; t_1 < t_2.$$

Year	Frontier	Output levels X in ktonnes							
		50	100	150	200	250	300	350	400
1920/29	1.49	5.19	2.83	2.34					
1929/32	1.56	1.56	1.61	1.79	1.82	1.64	1.31	1.07	
1932/33	1.19	1.11	1.14	0.98	0.88	1.12	1.10	1.04	
1933/37	1.26	1.17	1.50	1.68	1.73	1.32	1.36	1.48	
1937/43	0.97	0.89	0.61	0.57	0.55	0.68	1.16	0.97	
1943/46	1.07	1.35	1.22	1.13	1.10	1.04	0.70	0.83	
1946/51	0.35	0.58	0.79	0.90	1.05	1.00	0.92	0.98	
1951/52	0.88	1.01	1.00	0.96	0.91	0.95	1.01	0.96	
1952/54	1.81	1.22	1.22	1.22	1.23	1.20	1.17	1.14	
1954/74	17.98	14.13	12.46	12.49	11.30	10.76	10.31	10.33	9.68
1920/74	36.88	127.47	69.34	58.46					

9.5 Concluding remarks

The development of the Swedish pulp industry in the 50-year period shows three distinct phases. During the prewar years overall cost reducing technical progress was fairly rapid and energy was substituted for labour. During the war period and the Korea-boom years this development was reversed. Costs were increasing and labour substituted for energy. Then during the last 20 years, 1954–74, we have a period of substantial substitution of energy for labour together with cost reductions.

The impact of energy-labour substitution reversals are clearly brought out in the short-run function diagrams. The substitution region for 1954 swings back toward the labour axis particularly for sulphite and sulphate pulp. The capacity region figures reveal that for mechanical pulp there has been a marked long-run substitution process between energy and labour, but with the cost reductions due to labour saving far outweighing cost increases due to higher energy input coefficients. The capacity region for sulphite and sulphate pulp indicate a long-run technical change reducing both types of input coefficients, i.e., a simultaneous productivity improvement of labour and energy.

10

Swedish Pig Iron Production

10.1 Introduction

The purpose of this chapter is to analyse long-run technical change by utilising the short-run industry production function for Swedish blast furnaces producing pig iron. The time span is rather long — the 6 cross-section samples cover the years 1850, 1870, 1913, 1935, 1950 and 1975.

10.2 Data

The data were originally collected by Wibe [1980]. For the years 1850–1913 they were extracted from primary data at the Statistics Sweden and the Swedish Iron Association. Because pig iron production has been important for Swedish exports during the periods in question, statistics recorded for this industry have been of high quality. The data for the period 1935–50 are based on primary data from Statistics Sweden; for the year 1975 the data have been collected directly from the firms in question. The data cover 90–95 per cent of the total production. Total production has increased from 100,000 tonnes in 1850 to 3.5 million tonnes in 1975.

Three basic techniques were employed during the period 1850–1974: charcoal, electric and later coke blast furnaces. Charcoal was the only technique in 1850 and 1880, while all three techniques were in use in 1913, 1935 and 1950. Only coke furnaces were in use by 1975.

The assumption of fixed-input coefficients *ex post* is very appropriate for blast furnaces. The only question in this connection is whether one year is too long a period for the assumption to hold. Although scrapping and day-to-day improvements occur continuously, concentrating on the units existing at the end of each of the six chosen years and assuming fixed

coefficients for each of these years should constitute a most satisfactory approximation.

The following notation will be employed:

L = labour (hours)
E = energy (Gcals)
X = output (tonnes)
L/X, E/X = input coefficients.

10.3 The short-run industry production function

The development of the short-run industry function is shown in Figure 10.1. The charcoal technique was more or less the same between 1850 and 1880. The only change that took place was that the average performance moved towards the stationary best practice performance.

From 1880 onwards the substitution region has shifted steadily towards the energy axis. In 1913 the two new techniques, electric and coke furnaces, were introduced. The average size of the charcoal units also increased. The substitution region is at its widest in the years 1913, 1935, and 1950, when the three processes were in use at the same time.

The productivity improvement can be seen by following an isoquant representing the same output level from period to period. The distance between the isoquants is 50 ktonnes for 1850–1935, and 100 ktonnes for 1950 and 1975. The movement of the isoquant for 300 ktonnes of pig iron is indicated by the arrows in Figure 10.1.

The shape of the isoquants is on the whole about the same except for the last period. The labour substitution possibilities are the greater, measured on a per centage basis. This is natural, since both a substitution effect and a technical change effect act simultaneously to reduce the labour coefficients. In 1975 all the units were very similar, especially with respect to labour coefficients. The greatest scope for short-run substitution was provided in the energy dimension. This development is clearly portrayed in Figure 10.2, where the capacity regions for 1850 and 1975 are shown.

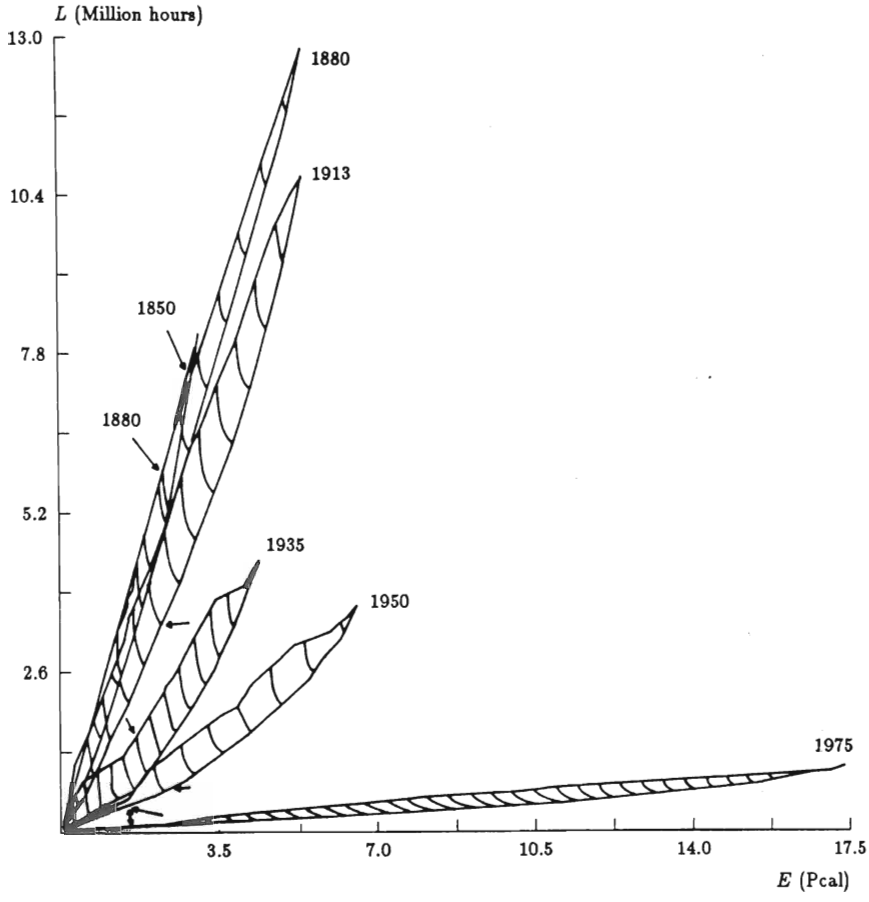


Figure 10.1: Isoquant maps of the short-run industry functions 1850–1975 for pig iron production.

10.4 A case of coexisting production techniques

To investigate more closely the utilisation of units with different technologies we have chosen to look at 1935. The capacity distribution for this year is shown in Figure 10.3. The two newest techniques, electric furnaces and coke ovens, have distinctively different characteristics with electric furnaces the most energy efficient and coke ovens the most labour efficient.

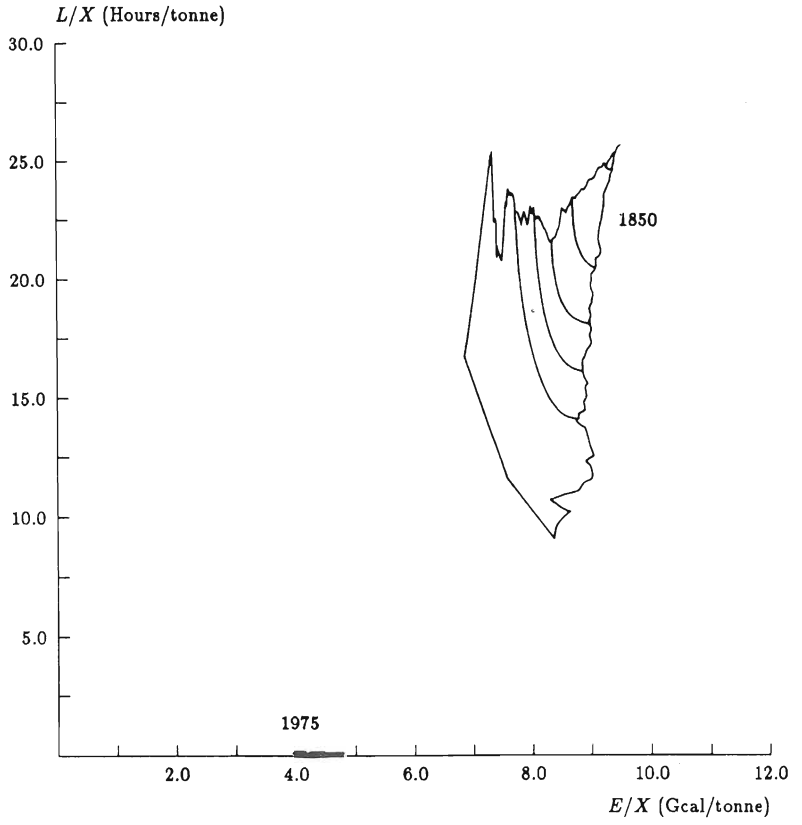
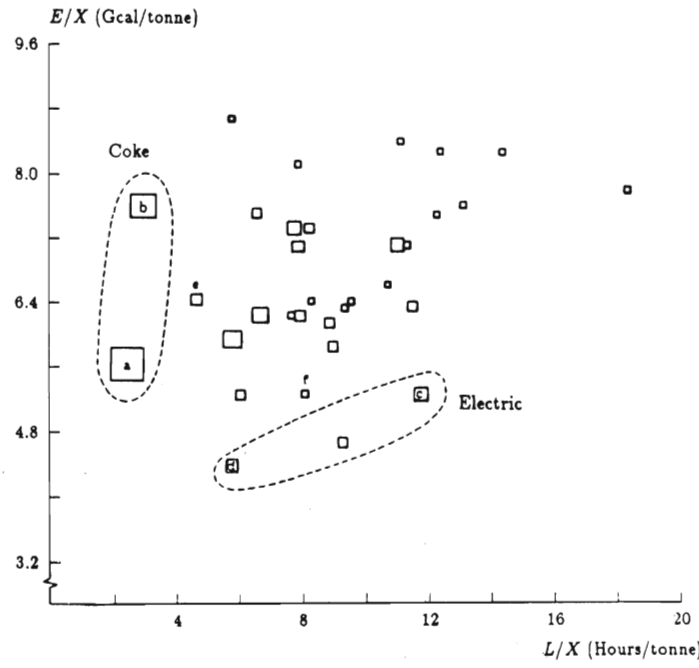


Figure 10.2: The development of the capacity region 1850–1975 for pig iron production.

The newest coke ovens are significantly larger than their predecessors, explaining their low labour-input coefficients. The input coefficients of the generally smaller charcoal ovens are distributed over a wide interval in both dimensions. A complete representation of the utilisation of units is shown in Figure 10.4.[1]

[1] For an elaborate treatment of this presentation technique, see Section 5.3 and Appendix 5.1.



The size of the squares are proportional to capacity.

Figure 10.3: The capacity distribution of Swedish blast furnaces in 1935.

The two coke ovens (a,b) are the first to be utilised along the upper boundary. While the most labour efficient unit (a) is fully utilised at an early stage of industry production, the least labour efficient oven (b) is not utilised fully independent of relative factor prices,until the entire capacity of the industry is almost exhausted. The utilisation pattern of the largest electric furnace (and least efficient in both dimensions[2]), electric furnace (c), is even more remarkable. Starting near the origin at the lower boundary (i.e., a relatively high energy efficiency) the utilisation strip goes through the substitution region almost to the exhaustion of total capacity at the upper boundary.

The substitution region is divided into two parts with different average

[2] See Figure 10.3.

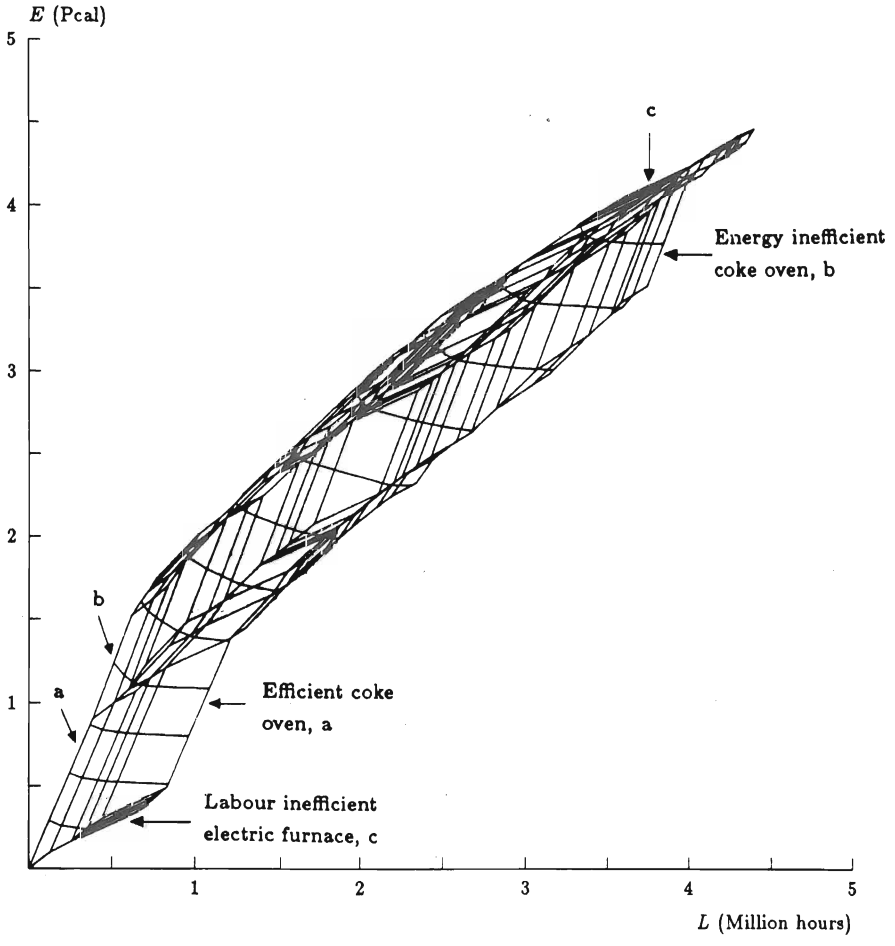


Figure 10.4: The region of substitution and the utilisation pattern of micro units in the short-run industry production function, 1935.

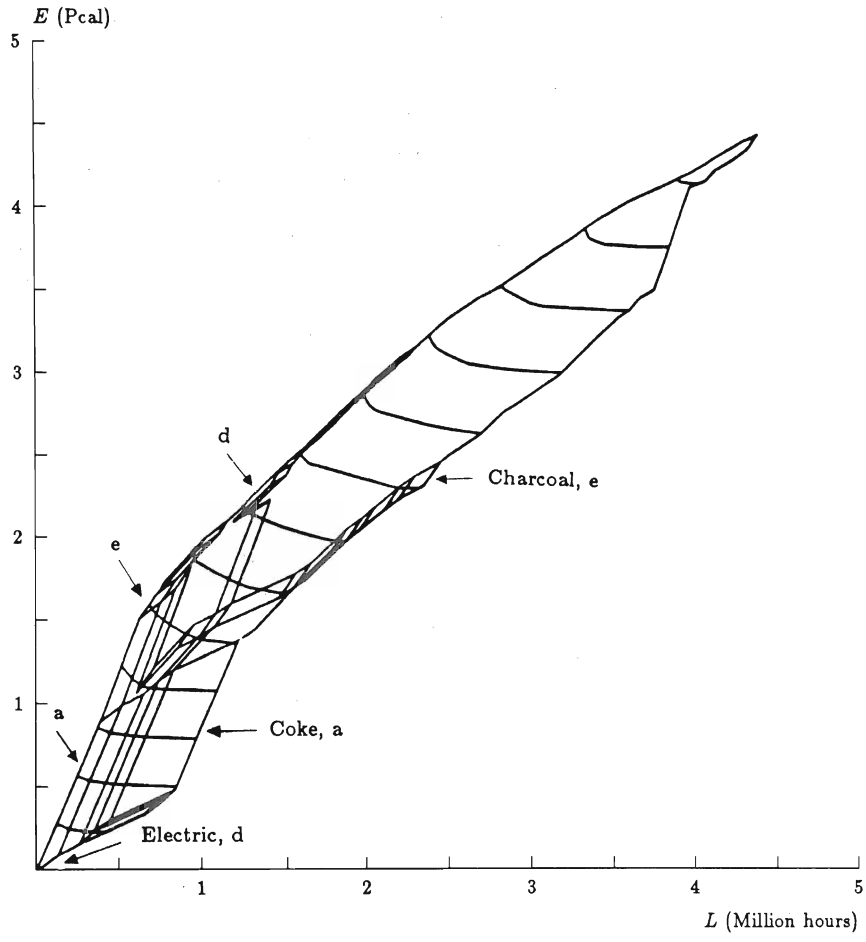


Figure 10.5: The utilisation pattern of selected micro units in the short-run industry production function, 1935.

labour-energy proportions. The kink in the region occurs just where the first coke unit is fully utilised at the lower boundary. The first part of the substitution region is made up of the electric furnace units, the coke units and some of the efficient charcoal units. As can be seen from the capacity diagram, Figure 10.3, these units have on the average a lower labour-energy

ratio. The utilisation patterns of the most efficient units of each technique, one electric (d), one coke (a), and two charcoal, (e) and (f), are illustrated in Figure 10.5.

The utilisation strip of the most efficient unit, electric furnace (d), starts at the origin and moves along the upper boundary. The utilisation strip of the most energy efficient charcoal unit, (f), also starts fairly close to the origin; it then moves across the substitution region into the region's interior, moving within the interior until it finally reaches the upper boundary at a fairly high level of industry capacity utilisation. The utilisation strip for the most labour efficient charcoal unit, (e), moves across the substitution region in a V-pattern.

Some isoquants are also shown in Figure 10.5. Generally, the scope for labour substitution is greater than for energy.

10.5 Technical progress

In this section we look more closely at the process of technical change on the basis of Salter's measures of technical advance and bias change.

Numerical measures of technical advance for selected output levels are set out in Table 10.1. In the period 1850–80 there was very little technological development. Technical advance results from average practice catching up with best practice. For all the other periods there is a steady technical advance, particularly strong in the period 1913–35 when new technologies were introduced and in the postwar period. A special feature of this period is the significantly larger technical advance the higher the output level, the difference in technical advance between the frontier and total industry production being about 20 per centage points. This period is characterised by a marked decrease in the number of units and investment in larger furnaces. The overall technical advance for the 125-year period has been about 70 to 85 per cent in cost reductions calculated in 1975 prices.

The nature of technical change is revealed in Table 10.2. The concentration of the capacity region towards best practice in the period 1850–80 shows up as both labour-using change and labour-saving change, depending on output level. But for all other periods there is a uniform labour-saving bias. This is particularly strong in the postwar years and for higher output levels. The much larger furnaces coming on stream in this period made strong labour-saving bias possible. Over the 125-year period the optimal energy/labour ratio increased from a factor of about 20 to about 50.

Table 10.1: The Salter technical advance measure T in 1975 prices.

$$T = \frac{C_{t_1}}{C_t}\Big|_{X=X^0}, \; C_t = \text{ minimised cost in year } t.$$

Year	Frontier	Output levels X in 10 ktonnes					
		5	10	15	20	25	30
1850/1880	1.17	1.00	0.97	0.94	0.90	0.85	0.79
1880/1913	0.72	0.71	0.71	0.70	0.69	0.70	0.70
1913/1935	0.61	0.58	0.56	0.56	0.57	0.59	0.60
1935/1950	0.79	0.79	0.79	0.79	0.76	0.72	0.72
1950/1975	0.69	0.69	0.69	0.69	0.69	0.69	0.67
1850/1975	0.28	0.22	0.21	0.20	0.19	0.17	0.16

Year	Output levels X in 10 ktonnes						
	35	40	45	50	55	60	65
1850/1880							
1880/1913	0.70	0.71	0.71	0.71	0.72	0.72	
1913/1935	0.61	0.61	0.63	0.64	0.64	0.65	0.66
1935/1950	0.73	0.73	0.73	0.72	0.71	0.70	0.70
1950/1975	0.64	0.62	0.60	0.58	0.57	0.57	0.56
1850/1975							

Year	Output levels X in 10 ktonnes						
	70	75	80	85	90	95	100
1850/1880							
1880/1913							
1913/1935							
1935/1950							
1950/1975	0.55	0.54	0.54	0.53	0.52	0.51	0.50
1850/1975							

Table 10.2: The Salter factor bias measure D_{EL} in 1975 prices.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}} \frac{L_{t_1}}{L_{t_2}}\Big|_{X=X^0}, \; t_1 < t_2.$$

Year	Frontier	Output levels X in 10 ktonnes					
		5	10	15	20	25	30
1850/1880	0.59	0.86	0.95	0.96	1.00	1.04	1.12
1880/1913	2.05	1.72	2.19	2.49	2.19	2.01	1.87
1913/1935	2.04	2.48	2.01	1.87	1.66	1.94	2.25
1935/1950	1.54	1.54	1.54	1.54	2.02	1.94	1.78
1950/1975	5.54	5.54	5.54	5.54	5.54	5.54	5.76
1850/1975	21.01	31.37	35.46	37.95	40.87	43.45	48.22

Year	Output levels X in 10 ktonnes						
	35	40	45	50	55	60	65
1850/1880							
1880/1913	1.80	1.73	1.69	1.61	1.57	1.56	
1913/1935	2.09	1.93	1.82	1.81	1.79	1.68	1.63
1935/1950	1.77	1.86	1.93	1.96	1.97	2.10	2.32
1950/1975	6.60	7.13	7.50	7.84	8.14	8.08	7.08
1850/1975							

Year	Output levels X in 10 ktonnes						
	70	75	80	85	90	95	100
1850/1880							
1880/1913							
1913/1935							
1935/1950							
1950/1975	6.32	6.51	6.56	6.68	6.88	7.27	7.48
1850/1975							

10.6 Concluding remarks

In this chapter we have briefly analysed the development of Swedish pig iron production during an extremely long time period. During the entire period we have found a gradual reduction in unit costs and also a labour-saving bias, except for some output levels during the first subperiod 1850–80. The labour-saving bias is more pronounced at higher output levels and the same holds for unit-cost reductions. When looking at the entire period though it varies somewhat in the different subperiods. In this chapter we have also utilised the activity region representation, introduced in Section 5.3, to illustrate the use of different technologies in the substitution region.

11

The Norwegian Aluminium Industry*

11.1 Introduction

The Norwegian aluminium industry is extremely electricity intensive, accounting for about twelve per cent of the total electricity consumption in Norway during 1978. The fact that the aluminium plants are located in remote regions means the industry is an important source of regional employment. Nevertheless, there is a lively debate within Norwegian society about the social profitability of the aluminium industry since it is charged less for its electricity consumption than the average electricity price level would dictate.*

In this chapter we establish short-run industry production functions of the Norwegian primary aluminium industry in order to analyse the technical progress and structural change that occurred in this industry during the period 1966–1978. We hope this analysis will contribute to a better understanding of the forces which have underpinned the restructuring of the aluminium industry in Norway.

There are good technical reasons for accepting the clay half of our putty-clay assumption as an appropriate and realistic assumption for this particular industry.[1] Our unit of observation is the plant producing primary aluminium by electrolysis from raw aluminium. Each plant might contain quite a few different vintages of smelters. (The ideal unit for our approach would have been the smelter itself.) Engineering information reveals that raw aluminium can be considered as a shadow factor of production. Thus, we have restricted the current factors under study to labour and electricity.

---

* This chapter is based on sections first presented in Førsund and Jansen [1983a,b].

[1] See Johansen and Thonstad [1979].

11.2 Data and structural description

From the Norwegian Industrial Statistics we have had access to data on the Norwegian aluminium industry for the years 1966–78 .[2] The number of production units is relatively small, varying from 7 to 9 units during the period under study.

When describing the structural changes in the aluminium industry, we focus on four years spaced equally apart: 1966, 1970, 1974 and 1978. The notation employed is:

L = labour (hours)
E = electricity (kWh)
X = output (tonnes)
L/X, E/X = input coefficients, measured by observed inputs and outputs

The observed input coefficients for labour and energy for the years 1966 and 1978 are shown in Figure 11.1, which also reveals the change in the capacity distribution. The size of the squares is proportional to capacity. The production capacity has generally been increasing for every unit except the smallest ones. The relatively larger reduction in labour-input coefficients (versus energy-input coefficients) is clearly depicted by the almost horizontal shift of the capacity distribution.

With respect to the partial input-coefficient distributions (not shown here) we note that the shape of the labour input-coefficient distribution has changed significantly from an even, cumulative distribution to one with a constant level and a marked tail for approximately the last 10 per cent of industry capacity.

The downward movement over time, i.e., the uniform increase of labour productivity, is almost at a standstill between 1974 and 1978. The right-hand tail with its little productivity improvements consists of units with very small capacity shares.

Relatively speaking, the downward change over time of the energy input-coefficient distribution has been smaller than that for labour. There is, however, an obvious downward trend between 1966 and 1974, whereas the input coefficients are systematically higher in 1978 as compared to 1974 (except at the tails of the distributions). The distributions are all

[2] We have also gathered additional information about the capacity output in each plant, which is well-defined for this sector.

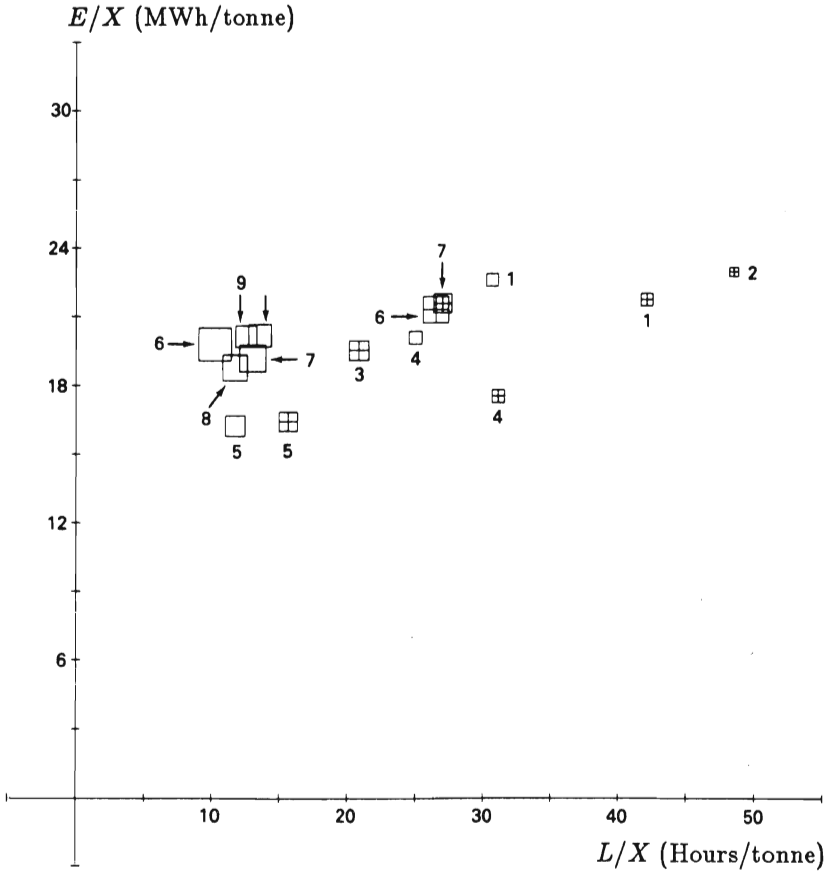


Figure 11.1: The capacity distributions in 1966 (crossed squares) and in 1978 (empty squares).

comparatively flat with tails for the last 5–10 per cent of industry capacity. The range of variation is from about 16,000 to 23,000 kWhs per tonne, excluding one extreme observation due to closure. The overall shift of the distributions amounts to a reduction of about 1,000 kWhs per tonne between 1966 and 1974, except for the stable 5–10 per cent tail.

Structural changes and the introduction of new production techniques are usually considered to be closely related to investment in new capital equipment. In the short run real capital may be considered as fixed, but

of course it also usually changes over time.

There has been a marked upward shift in the distribution of real capital per tonne aluminium over time. This should be an expected result of the vintage nature of the aluminium industry and of *a priori* knowledge about long-run substitution possibilities between the variable inputs labour and energy and capital. Moreover, we also find that the form of the capital-output distribution has changed over time in the same way as the labour-input coefficient, i.e., from an even cumulative distribution to one with a constant level and a marked tail for the last share of the industry capacity. The correlation across firms between the capital-output coefficients and the input coefficients of energy and labour, respectively, have changed considerably over time. Both correlation coefficients were clearly negative in 1966. In 1978, while there was no correlation between the energy coefficient and the capital-output ratio, there was a positive correlation between the labour-input coefficient and the capital-output ratio.

11.3 The short-run function and technical change

Information about the ex post micro production functions must be available in order to derive the short-run industry function. The production capacity of each unit is observed directly, and the fixed current-input coefficients are calculated by means of the observed amounts of current inputs and output. If the assumptions made about the ex post technology are valid, this is an appropriate procedure.

The region of substitution

The region of substitution and the isoquant map of the short-run industry production function for the selected years are shown in Figure 11.2. The substitution regions are rather narrow for all years, which is a reflection of the uniformity of the techniques utilised in Norwegian aluminium plants. The collapse of the substitution region into a single line, as is the case with the tail end in 1966, 1974 and 1978, and the front end in 1966, 1970 and 1974, corresponds to one unit obtaining the same rank number in the two partial input coefficient distributions.[3] The remaining scope for sub-

[3] The probability of this occurring is, of course, higher the smaller the number of production units. Recall that there are only between 7 and 9 units here.

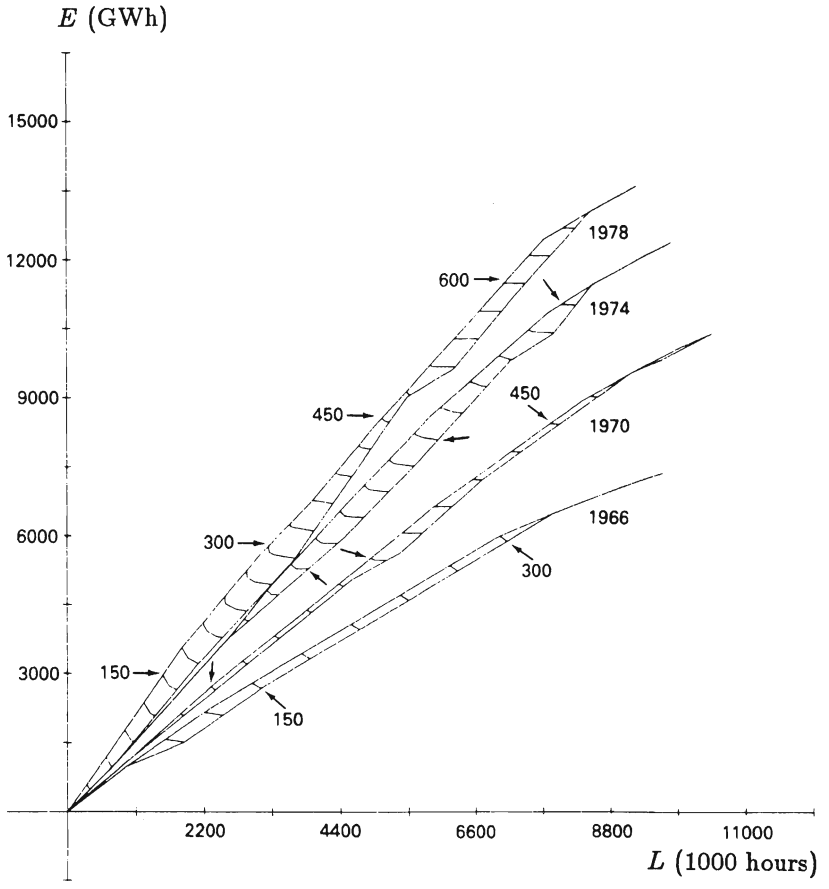


Figure 11.2: The development of the short-run industry production function between 1966 and 1978. The interval between the isoquants is 30,000 tonnes.

stitution at the industry level is markedly greater for labour, as should be expected from the structural description provided in the previous section.

This last observation is also valid as an explanation of the steady shift towards the energy axis revealed in Figure 11.2. In this context it should also be noted that there are strict physical limitations on the improvement in electricity productivity. According to Johansen and Thonstad [1979] there is within the existing technology very little feasible improvement left

of the best-practice electricity input-coefficient at the 1978 level, while the reduction in labour coefficients does not come up against any such physical law (except, of course, the level zero).

These shifts of the substitution region towards the energy axis are consistent with the changes observed in the relative input prices. The development of the prices shows, with a few exceptions, a steady increase in the price of labour relative to that of electricity, so that the relative price nearly doubles during the period of observation.

We note that the isoquants are almost straight lines with only a few corner points. Generally, the curvature of an isoquant is characterised by the elasticity of substitution. Short-run elasticities can be approximated by means of the analogy with the definition used in the case of smooth isoquants.[4] The change in the factor ratio relative to the average factor ratio, measured at the extreme points of two consecutive isoquant segments, is related to the change in the marginal rate of substitution between the two segments relative to the average rate of substitution. Contrary to the visual impression of the isoquants approximating straight lines, which would imply high values for the elasticity of substitution, we find rather low estimates of the elasticity of substitution between labour and electricity, a result that, nevertheless, corroborates the conjectures in Hildenbrand [1981].

The demand regions

What implications does the short-run industry production function have with respect to industry demand for inputs? A simple transformation of the substitution regions shown in Figure 11.2 yields the region within which the demand functions must lay for any set of input prices. Figures 11.3 and 11.4 show the demand regions for labour and electricity, respectively. The regions are projections of the borders of the substitution region in the three-dimensional space of two inputs and output into two-dimensional spaces of one input and output.

The upward shift of the labour demand regions is clearly noticeable. The demand regions for electricity are extremely narrow, ray-like, and stable over time.

Productivity changes

The productivity improvements for various levels of output can be studied in Figure 11.2 by following the shift between years of the isoquants in

[4] This has been shown in Section 5.4.

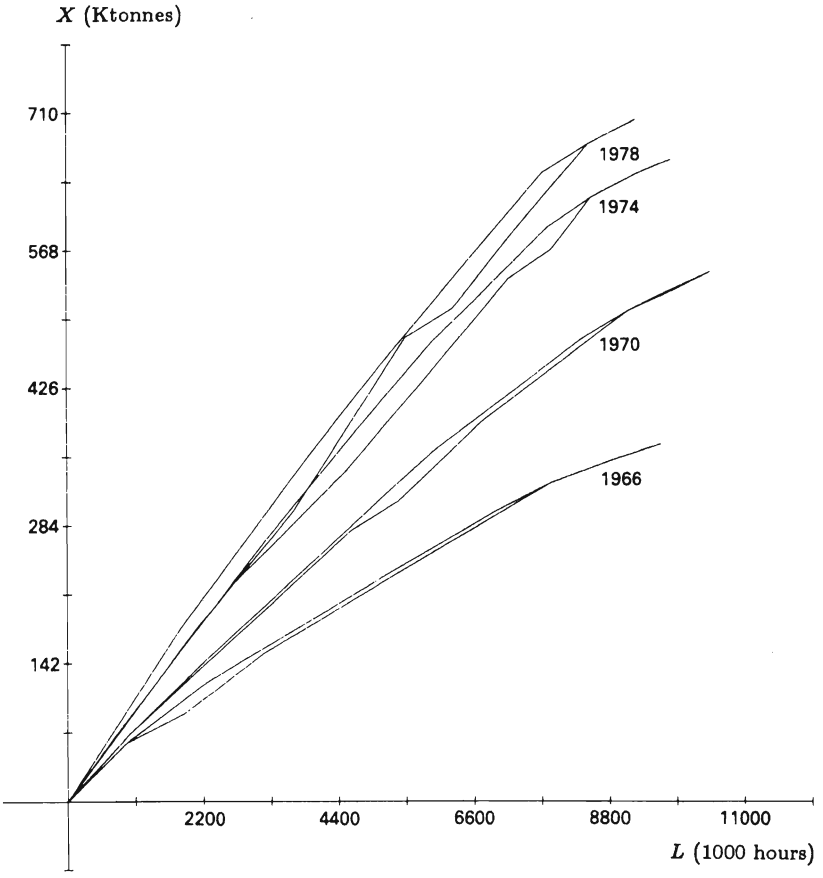


Figure 11.3: The demand region for labour for the years 1966, 1970, 1974 and 1978.

question. The interval length in Figure 11.2 is 30 ktonnes. The levels of 150 ktonnes, 300 ktonnes, 450 ktonnes and 600 ktonnes are shown separately in Figure 11.5. The almost exclusively labour-saving movement is clearly portrayed. Energy productivity has, as a matter of fact, decreased due to the shift from the high capacity utilisation of 1974 to the lower rate of capacity utilisation in 1978 for all levels of output.

The movement towards the electricity axis is also clearly revealed by the isoquant maps within the substitution regions. These have been trans-

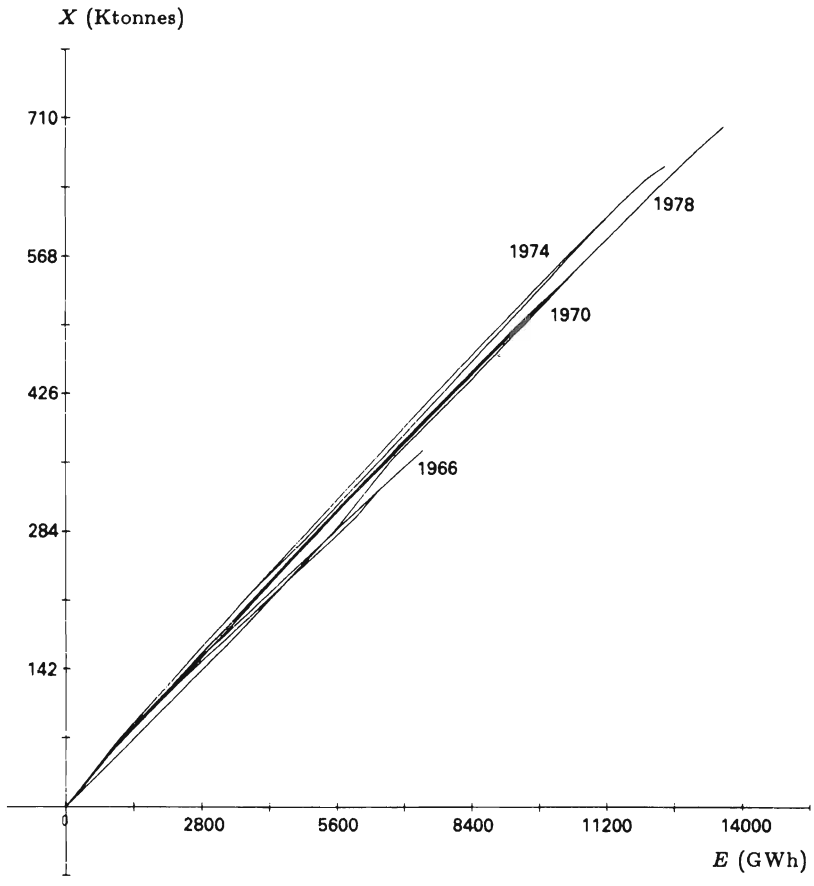


Figure 11.4: The demand region for electricity for the years 1966, 1970, 1974 and 1978.

formed from the input space of Figure 11.2 into the input-coefficient space in Figure 11.6. The transformations represent the feasible regions of the input coefficients for the short-run industry function, and must therefore show more limited variations than the capacity distributions of individual units shown in Figure 11.1.

As far as energy usage is concerned, Figure 11.6 shows that the frontier values of the electricity input-coefficients have been quite stable except for one particular unit in 1974. The industry improvement has consisted of the

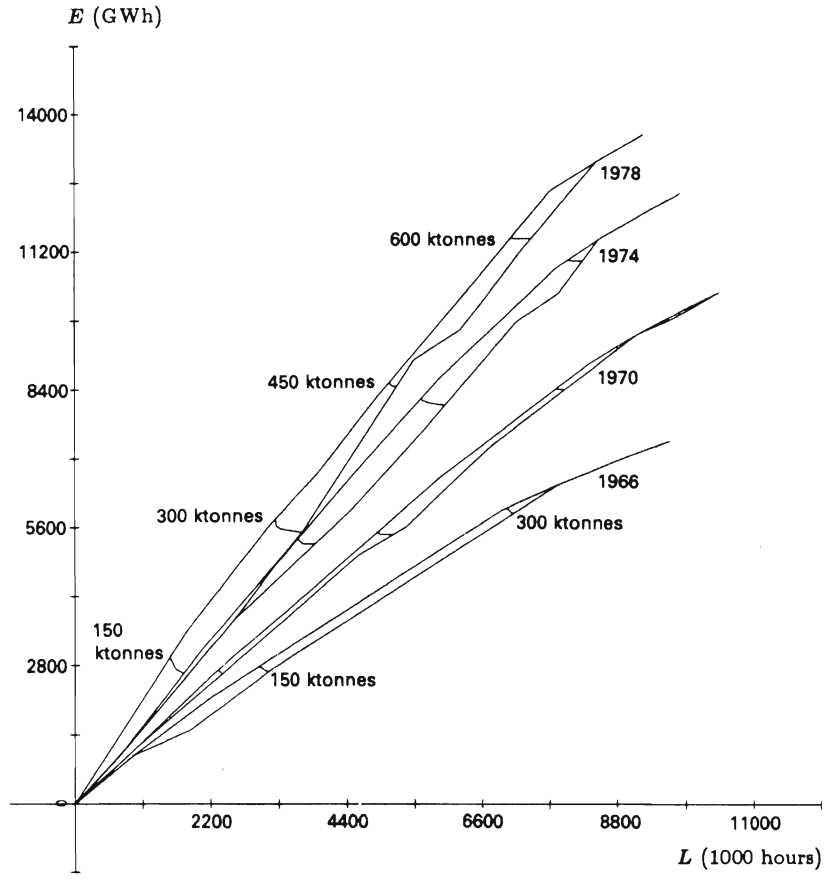


Figure 11.5: The short-run industry production functions for the years 1966, 1970, 1974 and 1978 with the isoquants of 150, 300, 450 and 600 ktonnes.

other units catching up with best-practice performance. This trend weakened between 1974, the year of high capacity utilisation, and 1978, a year with a less than average rate of capacity utilisation. The movements of the isoquants over time are more sharply brought out by their transformation to the input-coefficient space. The movement towards the southwest up to 1974, and the increase in electricity coefficients in 1978 to about the same level as in 1970 are clearly visible here.

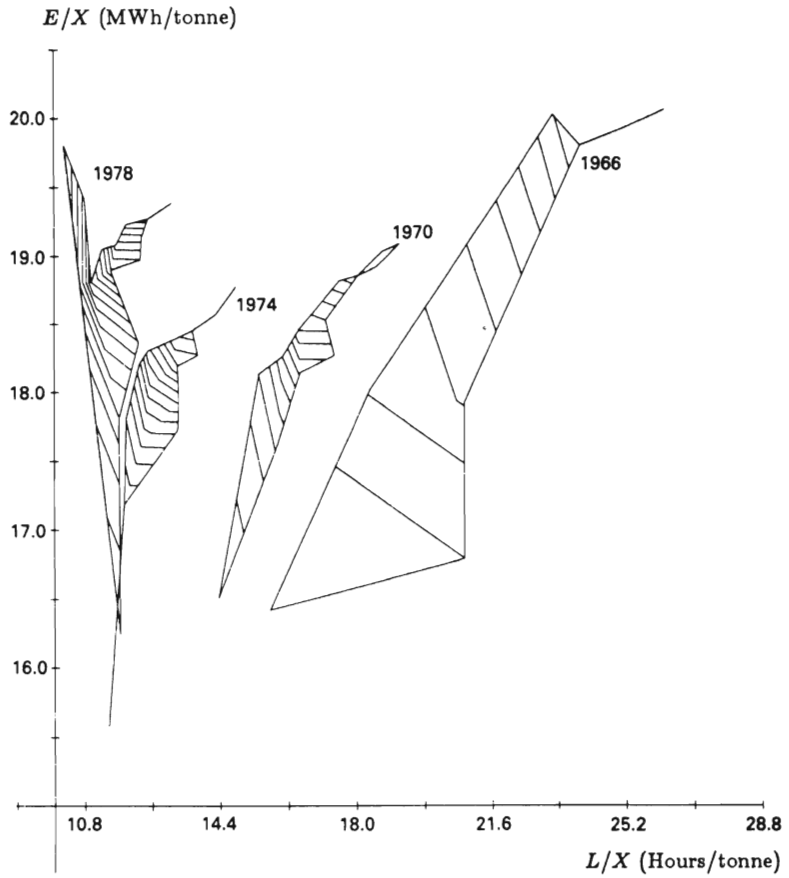


Figure 11.6: The development of the capacity region of the short-run industry production function for the years 1966, 1970, 1974 and 1978.

Measures of technical progress

As discussed in Section 3.6, the significance of technical change can be assessed by computing the relative change in unit costs at constant input prices and output levels. We have chosen to use the average observed prices in the last sample year, 1978. The results for output intervals of

Table 11.1: The Salter technical advance measure T in 1978 prices.

$$T = \frac{C_{t+1}}{C_t}\Big|_{X=X^0}, \quad C_t = \text{minimised unit cost in year } t.$$

Year	Frontier	Output levels, X, in ktonnes			
		150	300	450	600
1966/70	0.95	0.86	0.78		
1970/74	0.86	0.84	0.85	0.83	
1974/78	1.04	1.01	0.98	0.98	0.96
1966/78	0.85	0.72	0.65		

150 ktonnes, including the frontier, i.e., the best-practice performance, are shown in Table 11.1.

The unit-cost reduction between 1966 and 1978 varied significantly: from the frontier, which shows a reduction of about 15 percent, to a much higher reduction of unit costs at higher output level, e.g., 35 per cent at 300 ktonnes. Corresponding to what was revealed by Figure 11.6, the only significant improvement of the frontier was between 1970 and 1974, but this was due to just one individual unit, and the performance slipped again, resulting in an *increase* of unit costs at best-practice between 1974 and 1978. The average catching up with best-practice performance shows up in Table 11.1, where the greatest unit cost reductions are shown to occur at higher output levels. The technical advance between 1974 and 1978 was very small indeed, the reduction in labour-input coefficients barely offsetting *increases* in electricity-input coefficients. Technical progress is here measured in terms of reductions in current costs. To complete the picture, capital costs should, of course, also be taken into consideration.

The factor-saving bias is expressed by computing the Salter [1980] measure of bias, i.e., the change in the cost-minimising factor ratios for consecutive time periods, keeping factor prices constant.

At the frontier the electricity-labour ratio increased by 32 per cent for the entire period, the most significant change taking place between 1966 and 1974. The labour-saving bias is greater, the higher the output level, and is 102 per cent for the entire period at the output level of 300,000 tonnes.

Table 11.2: The Salter factor bias measure D_{EL} in 1978 prices.

$$D_{EL} = \frac{E_{t_2}}{E_{t_1}} \frac{L_{t_1}}{L_{t_2}}\Big|_{X=X^0}, \; t_1 < t_2.$$

Year	Frontier	Output levels, X, in ktonnes			
		150	300	450	600
1966/70	1.10	1.25	1.31		
1970/74	1.19	1.22	1.32	1.36	
1974/78	1.01	1.16	1.16	1.13	1.18
1966/78	1.32	1.77	2.02		

Only a few points on the average cost curves were used in Table 11.1. The complete average cost curves for 1966, 1970 and 1978 are shown in Figure 11.7 together with the marginal cost curves. All are based on the 1978 average observed input prices.[5]

Salter measures at various output levels may be calculated by comparing average costs in Figure 11.7. The average cost curve has flattened out noticeably.

The shape of the marginal cost curves add to the structural picture. They have become more and more like the average cost curves, with the tails of the J-shapes applicable to smaller and smaller shares of output capacity. This development supports the impression of an increasing uniformity of the structure to aluminum smelters.

The elasticity of scale

Additional structural features can be identified by studying the values of the elasticity of scale. In Table 11.3 the development of the scale elasticity is shown for the average factor ratio. When the factor ray is outside the substitution region, the scale elasticity along the bordering isoquant segment in question is used.

The maximal value of the scale elasticity in short-run industry functions is 1.0. The level of the elasticities has increased between 1966 and

[5] The curves for 1974 are excluded because of their proximity to the 1978 curves, as is evident from Table 11.1.

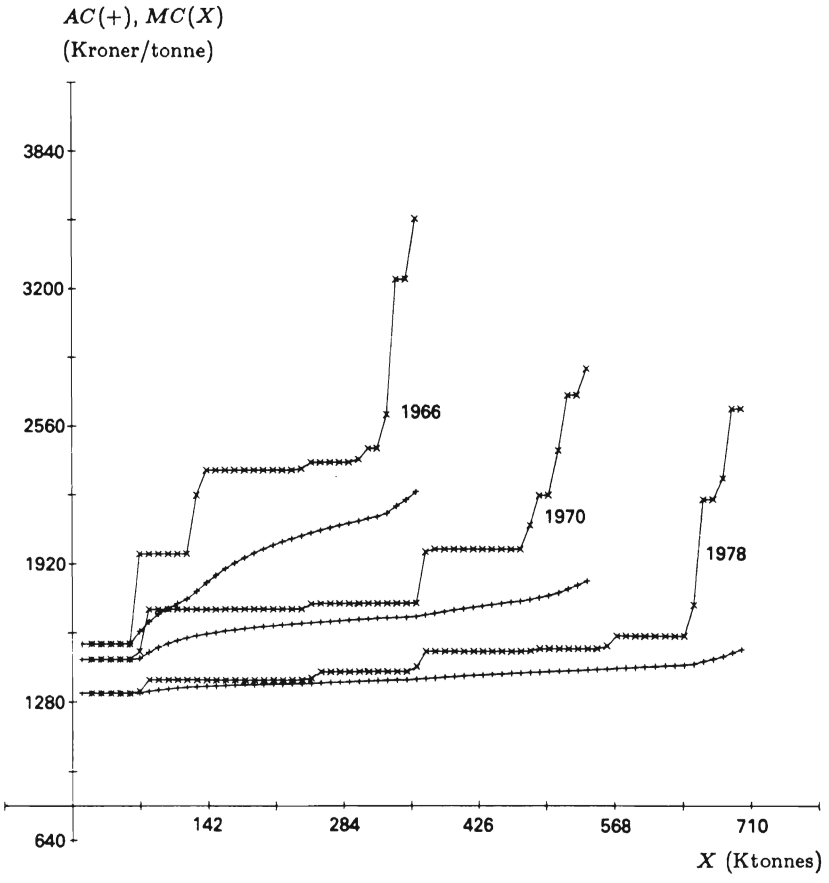


Figure 11.7: The average and marginal cost-functions (AC and MC, respectively) for 1966, 1970 and 1978 in 1978-prices.

1974. The high values in 1974 and 1978 again reflect the technical uniformity of the units. The extremely low value for the highest output level in 1978 is due to the fact that the least efficient unit was then utilised, the latter unit corresponding to the top of the tail of the J-shaped marginal cost curve for that year.

Table 11.3: The development of the scale elasticity along the average factor rays.

Year	Output levels in ktonnes							Energy/ labour *
	100	200	300	400	500	600	700	
1966	0.89	0.86	0.89					0.76
1970	0.91	0.94	0.94	0.92	0.94			1.00
1974	0.96	0.98	0.94	0.96	0.95	0.91		1.26
1978	0.94	0.93	0.95	0.97	0.94	0.95	0.42	1.47

* Average factor ratio

11.4 Concluding remarks

There has been a marked shift of the substitution region towards the electricity axis. Direct substitution between electricity and labour is possible only to a very limited extent when capital is a variable factor. Thus we interpret the above results as clear evidence of labour-saving technical change during the period of observation. This change has probably been induced by the rise in the relative price of labour, by 200 per cent between 1966 and 1978, while another factor has been the increased technical possibilities for cost reduction. Assuming that this process continues, the regional employment impact of this industry will lessen.

The short-run industry production function for aluminium is characterised by narrow substitution regions for all years, reflecting a high degree of technical uniformity among Norwegian aluminum smelters.

This uniformity is partially a result of labour-saving investments undertaken by all plants at more or less the same time, but can also be seen as a result of small improvements in the basic process of smelting aluminium. The structure is therefore quite similar to the one that appears in long-run steady state with no technological change. From an economic policy point of view the structure of 1978 implies that the *entire* industry could run into deficit during a period of falling aluminum prices on the world market.

From a regional point of view the above observations mean that employment in the aluminum industry is extremely vulnerable to fluctuations in world market prices for aluminium.

References

Afriat, S.N. [1972], "Efficiency Estimation of Production Functions", *International Economic Review, 13*, 568–98.

Aigner, D.J., T. Amemiya and D.J. Poirier [1976], "On the Estimation of Production Frontiers: Maximum Likelihood Estimation of the Parameters of a Discontinuous Density Function", *International Economic Review, 17*, 377–96.

Aigner, D.J. and S.F. Chu [1968], "On Estimating the Industry Production Function", *American Economic Review, 58*, 226–39.

Aigner, D.J., C.A.K. Lovell and P. Schmidt [1977], "Formulation and Estimation of Stochastic Frontier Production Function Models", *Journal of Econometrics, 6*, 21–37.

Albrecht, J.W. and A.G. Hart [1983], "A Putty-Clay Model of Demand Uncertainty and Investment", *Scandinavian Journal of Economics, 85*, 393–402.

Binswanger, H.P. [1974], "The Measurements of Technical Change Biases with Many Factors of Production", *American Economic Review, 64*, 964–976.

Bliss, C. [1968], "On Putty-Clay", *Review of Economic Studies, 35*, 105–32.

Box, G.E.P. and D.R. Cox [1964], "An Analysis of Transformation", *Journal of the Royal Statistical Society, 26*, 211–243.

Broeck, J. van den, F.R. Førsund, L. Hjalmarsson and W. Meeusen [1980], "On the Estimation of Deterministic and Stochastic Frontier Production Functions", *Journal of Econometrics, 13*, 117–38.

Carlsson, B. [1968], "The Measurement of Efficiency in Production: An Application to Swedish Manufacturing Industries 1968", *Swedish Journal of Economics, 74*, 468–85.

Carlsson, B. [1978], "Choice of Technology in the Cement Industry — A Comparison of the United States and Sweden", in *The Importance of Technology and the Permanence of Structure in Industrial Growth*, ed. B. Carlsson, G. Eliasson and I. Nadiri. Stockholm: The Industrial Institute for Economic and Social Research.

Charnes, A., W.W. Cooper and E. Rhodes [1978], "Measuring the Efficiency of Decision-Making Units", *European Journal of Operational Research, 2*, 429–44.

Charnes, A., W.W. Cooper and E. Rhodes [1981], "Evaluating Program and Managerial Efficiency: An Application of Data Envelopment Analysis to Program Follow Through", *Management Sciences, 27*, 668–97.

Ching, C.T.K. [1973], "A Note on the Stability of Firm Size Distribution Functions for Western Cattle Ranches", *American Journal of Agricultural Economics, 55*, 500–2.

Christensen, L.R. and W.H. Greene [1976], "Economies of Scale in U.S. Electric Power Generation", *Journal of Political Economy, 84*, 655–76.

Chu, S.-F. [1978], "On the Statistical Estimation of Parametric Frontier Production Functions: A Reply and Further Comments", *Review of Economics and Statistics, 60*, 479–81.

Danø, S. [1966], *Industrial Production Models*. Wien: Springer-Verlag.

Dhrymes, P.J. and M. Kurz [1964], "Technology and Scale in Electricity Generation", *Econometrica, 22*, 287–315.

Eide, E. [1979], *Engineering Production and Cost Functions for Tankers*. Amsterdam: Elsevier Scientific.

Elliot, J.E. [1980], "Marx and Schumpeter on Capitalism's Creative Destruction: A Comparative Restatement", *Quarterly Journal of Economics, 95*, 45–68.

Engwall, L. [1972], "Inequality of Firm Sizes in Different Economic Systems", *Zeitschrift für Nationalökonomie, 32*, 449–60.

Färe, R. and C.A.K. Lovell [1978], "Measuring the Technical Efficiency of Production: Reply", *Journal of Economic Theory, 19*, 150–62.

Färe, R. and C.A.K. Lovell [1981], "Measuring the Technical Efficiency of Production: Reply", *Journal of Economic Theory, 25*, 453–4.

Färe, R., S. Grosskopf and C.A.K. Lovell [1983], "The Structure of Technical Efficiency", *Scandinavian Journal of Economics, 85*, 181–90.

Färe, R., S. Grosskopf and C.A.K. Lovell [1985], *The Measurement of Efficiency of Production*. Boston: Kluwer-Nijhoff.

Farrell, M.J. [1957], "The Measurement of Productive Efficiency", *Journal of the Royal Statistical Society, 120*, 449–60.

Farrell, M.J. and M. Fieldhouse [1962], "Estimating Efficient Production Functions under Increasing Returns to Scale", *Journal of the Royal Statistical Society, 125*, 252–67.

Førsund, F.R. [1971], "A Note on the Technically Optimal Scale in Inhomogeneous Production Functions", *Swedish Journal of Economics, 73*, 225–40.

Førsund, F.R. [1974], "Studies in the Neoclassical Theory of Production", Memorandum from the Institute of Economics, University of Oslo, February 4.

Førsund, F.R. [1975], "The Homothetic Production Function", *Swedish Journal of Economics, 77*, 234–44.

Førsund, F.R. [1985–86], "Comment on Frontier Production Functions", *Econometric Reviews, 4*, 329–34.

Førsund, F.R. and L. Hjalmarsson [1974a], "On the Measurement of Productive Efficiency", *Swedish Journal of Economics, 76*, 141–54.

Førsund, F.R. and L. Hjalmarsson [1974b], "Comment on Bo Carlsson's 'The Measurement of Efficiency in Production: An Application to Swedish Manufacturing Industries 1968", *Swedish Journal of Economics, 76*, 251–54.

Førsund, F.R. and L. Hjalmarsson [1978a], "Technical Progress and Structural Efficiency of Swedish Dairy Plants", *Le capital dans la fonction de production*, Institut de recherche en économie de la production, Paris X-Nanterre, 101–21.

Førsund, F.R. and L. Hjalmarsson [1978b], "Production Functions in the Swedish Particle Board Industry", *Le capital dans la fonction de production*, Institut de recherche en économie de la production, Paris X-Nanterre, 79–99.

Førsund, F.R. and L. Hjalmarsson [1979a], "Frontier Production Functions and Technical Progress: A Study of General Milk Processing in Swedish Dairy Plants", *Econometrica, 47*, 883–900.

Førsund, F.R. and L. Hjalmarsson [1979b], "Generalised Farrell Measures of Efficiency: An Application to Milk Processing in Swedish Dairy Plants", *Economic Journal, 89*, 294–315.

Førsund, F.R. and L. Hjalmarsson [1983], "Technical Progress and Structural Change in the Swedish Cement Industry 1955–1979", *Econometrica, 51*, 1449–67.

Førsund, F.R. and E.S. Jansen [1977], "On Estimating Average and Best Practice Homothetic Production Functions via Cost Functions", *International Economic Review, 18*, 463–76.

Førsund, F.R. and E.S. Jansen [1983a], "Technical Progress and Structural Change in the Norwegian Primary Aluminum Industry", *Scandinavian Journal of Economics, 85*, 113–126.

Førsund, F.R. and E.S. Jansen [1983b], "Analysis of Energy-Intensive Industries — the Case of Norwegian Aluminium Production", in *Analysis of Supply and Demand of Electricity in the Norwegian Economy*, ed. O. Bjerkholt et al., Social Economic Studies, 53, Central Bureau of Statistics, Oslo.

Førsund, F.R. and E.S. Jansen [1983c], "The Interplay between Sectoral Models based on Micro Data and Models for the National Economy", Articles from the Central Bureau of Statistics, 142, Oslo.

Førsund, F.R. and E.S. Jansen [1985], "The interplay between sectoral models based on micro data and models for the national economy", in *Production, Multi-sectoral Growth and Planning*, ed. F.R. Førsund, M. Hoel and S. Longva. Amsterdam: North-Holland, 109–25.

Førsund, F.R., C.A.K. Lovell and P. Schmidt [1980], "A Survey of Frontier Production Functions and their Relationship to Efficiency Measurement", *Journal of Econometrics, 13*, 5–25.

Førsund, F.R., S. Gaunitz, L. Hjalmarsson and S. Wibe [1980], "Technical Progress and Structural Change in the Swedish Pulp Industry", in *The Economics of Technological Progress*, ed. T. Puu and S. Wibe. London: MacMillan.

Førsund, F.R., L. Hjalmarsson and Ø. Eitrheim [1985a], "An intercountry comparison of cement production: The short-run production function approach", in *Production, Multi-sectoral Growth and Planning*, ed. F.R. Førsund, M. Hoel and S. Longva. Amsterdam: North-Holland, 11–42.

Førsund, F.R., L. Hjalmarsson, J. Karko, Ø. Eitrheim and T. Summa [1985b], "An intercountry comparison of productivity and technical change in the Nordic cement industry", ETLA Report B44, Helsinki.

Freidenfelds, J. [1981], *Capacity Expansion Analysis of Simple Models with Applications*. Amsterdam: North-Holland.

Frisch, R. [1965], *Theory of Production*. Dordrecht: D. Reidel.

Fuss, M.A. [1977], "The Structure of Technology over Time: A Model for Testing the 'Putty-Clay' Hypothesis", *Econometrica, 45*, 1797–1822.

Fuss, M.A. and D. McFadden (eds.) [1978], *Production Economics: A Dual Approach to Theory and Applications*. Amsterdam: North-Holland.

Gabrielsen, A. [1975], "On Estimating Efficient Production Functions", Working Paper No. A-85, Christian Michelsen Institute, Department of Humanities and Social Sciences, Bergen.

Gibrat, R. [1957], "On Economic Inequalities", *International Economic Papers, 7*, 53–70.

Gilbert, J. and R.G. Harris [1984], "Competition with Lumpy Investment", *Rand Journal of Economics, 15*, 197–212.

Gould, J.P. [1968], "Adjustments in the Theory of the Firm", *Review of Economic Studies, 95*, 42–56.

Greene, W.H. [1980a], "Maximum Likelihood Estimation of Econometric Frontier Functions", *Journal of Econometrics, 13*, 27–56.

Greene, W.H. [1980b], "On the Estimation of a Flexible Frontier Production Model", *Journal of Econometrics, 13*, 101–115.

Greene, W.H. [1982], "Maximum Likelihood Estimation of Stochastic Frontier Production Models", *Journal of Econometrics, 18*, 285–9.

Greene, W.H. [1983], "Simultaneous Estimation of Factor Substitution, Economies of Scale, Productivity, and Non-Neutral Technological Change", in *Developments in Econometric Analyses of Productive Efficiency*, ed. A. Dogramaci. Boston: Kluwer-Nijhoff, 121–44.

Griliches, Z. and V. Ringstad [1971], *Economies of Scale and the Form of the Production Function*. Amsterdam: North-Holland.

Grosse, A. [1953], "The Technological Structure of the Cotton Textile Industry", in *Studies in the Structure of the American Economy*, ed. W. Leontief. New York: Oxford University Press.

Grosskopf, S. [1986], "The Role of the Reference Technology in Measuring Productive Efficiency", *Economic Journal, 96*, 499–513.

Haldi, J. and D. Whitcomb [1967], "Economies of Scale in Industrial Plants", *Journal of Political Economy, 75*, 373–85.

Heckscher, E.F. [1918], *Svenska produktionsproblem*. Stockholm: Bonniers.

Hildenbrand, K. [1983], "Numerical Computation of Short-Run Production Functions", in *Quantative Studies on Production and Prices*, ed. W. Eichhorn, R. Henn, K. Neumann and R. W. Shephard. Wien: Physica-Verlag, 173–80.

Hildenbrand, W. [1981], "Short-Run Production Functions based on Microdata", *Econometrica, 49*, 1095–1124.

Hjalmarsson, L. [1973], "Optimal Structural Change and Related Concepts", *Swedish Journal of Economics, 75*, 176–92.

Hjalmarsson, L. [1974], "The Size Distribution of Establishments and Firms Derived from an Optimal Process of Capacity Expansion", *European Economic Review, 5*, 123–40.

Hjalmarsson, L. [1975], "Studies in a Dynamic Theory of Production and its Applications", Department of Economics, University of Gothenburg, Memorandum No. 50.

Hjalmarsson, L. [1976a], "On Monopoly Welfare Gains, Scale Efficiency and the Costs of Decentralization", *Empirical Economics, 1*, 231–49.

Hjalmarsson, L. [1976b], "Reply", *European Economic Review, 7*, 287–92.

Ijiri, Y. and H.A. Simon [1964], "Business Firm Growth and Size", *American Economic Review, 54*, 77–89.

Ijiri, Y. and H.A. Simon [1971], "Effects of Mergers and Acquisitions on Business Firm Concentration", *Journal of Political Economy, 79*, 314–22.

Ijiri, Y. and H.A. Simon [1974], "Interpretations of Departures from the Pareto Curve Firm-Size Distributions", *Journal of Political Economy, 82*, 315–31.

International Dairy Federation [1974], "Optimum Size of Dairy Factories", General Secretariat, Brussels.

Johansen, L. [1959], "Substitution Versus Fixed Production Coefficients in the Theory of Economic Growth: A Synthesis", *Econometrica, 27*, 157–76.

Johansen, L. [1967], "Some Problems of Pricing and Optimal Choice of Factor Proportions in a Dynamic Setting", *Economica, 34*, 131–52.

Johansen, L. [1968], "Production Functions and the Concept of Capacity", in Recherches récentes sur la fonction de production, Economie mathématique et econometrie, 2. Namur: Ceruna, 47–72.

Johansen, L. [1972], *Production Functions*. Amsterdam: North-Holland.

Johansen, L. and Å Sørsveen [1967], "Notes on the Measurement of Real Capital in Relation to Economic Planning Models", *The Review of Income and Wealth, 13*, 175–97.

Johansen, L. and J. Thonstad [1979], "Aluminium Processing, Technology and Prospects", Final report (in Norwegian). SINTEF-rapport STF 34, Trondheim.

Jondrow, J., C.A.K. Lovell, I.S. Materov and P. Schmidt [1982], "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model", *Journal of Econometrics, 19*, 233–8.

Kemp, M.C. and P.C. Thanh [1966], "On a Class of Growth Models", *Econometrica, 34*, 257–82.

Komiya, R. [1962], "Technological Progress and the Production Function in the United States Steam Power Industry", *The Review of Economics and Statistics, 44*, 156–66.

Kon, Y. [1983], "Capital Input Choice under Price Uncertainty: A Putty-Clay Technology Case", *International Economic Review, 24*, 183–97.

Kopp, R.J. [1981a], "Measuring Technical Efficiency of Production: A Comment", *Journal of Economic Theory, 25*, 450–2.

Kopp, R.J. [1981b], "The Measurement of Productive Efficiency: A Reconsideration", *Quarterly Journal of Economics, 97*, 477–503.

Kopp, R.J. and W.E. Diewert [1982], "The Decomposition of Frontier Cost Function Deviations into Measures of Technical and Allocative Efficiency", *Journal of Econometrics, 19*, 319–31.

Kopp, R.J. and V.K. Smith [1980], "Frontier Production Function Estimates for Steam Electric Generation: A Comparative Analysis", *Southern Economic Journal, 47*, 1049–59.

Kopp, R.J. and V.K. Smith [1983], "Neoclassical Modeling of Non-neutral Technological Change: An Experimental Appraisal", *Scandinavian Journal of Economics, 85*, 127–46.

Kumbhakar, S. C. [1987], "The Specification of Technical and Allocative Inefficiency in Stochastic Production and Profit Frontiers", *Journal of Econometrics, 34*, 335–48.

Kurz, M. [1963], "Substitution vs. Fixed Production Coefficients: A Comment", *Econometrica, 31*, 209–17.

Lee, L.-F. [1983a], "A Test for Distributional Assumptions for the Stochastic Frontier Functions", *Journal of Econometrics, 22*, 245–67.

Lee, L.-F. [1983b], "On Maximum Likelihood Estimation of Stochastic Frontier Production Models", *Journal of Econometrics, 23*, 269–74.

Lee, L.-F. and W.G. Tyler [1978], "The Stochastic Frontier Production Function and Average Efficiency: An Empirical Analysis", *Journal of Econometrics, 7*, 385–89.

Lucas Jr., R.E. [1978], "On the Size Distribution of Business Firms", *Bell Journal of Economics, 9*, 508–23.

Lutz, F. and V. Lutz [1951], *The Theory of Investment of the Firm*. New York: Greenwood Press.

Manne, A.S. [1961], "Capacity Expansion and Probabilistic Growth", *Econometrica, 29*, 632–49.

Manne, A.S. (ed.) [1967], *Investments for Capacity Expansion*. London: Allen and Unwin.

Marshall, A. [1966], *Principles of Economics (Eigth Edition)*. London: MacMillan.

Marx, K. [1966], *Capital*. Moscow: Progress Publishers.

Maywald, K. [1957], "The Best and the Average in Productivity Studies and in Long-Term Forecasting", *The Productivity Measurement Review, 9*, 37–49.

McBride, M.E. [1981], "The Nature and Source of Economies of Scale in Cement Production", *Southern Economic Journal, 48*, 105–15.

Meeusen, W. and van den Broeck, J. [1977a], "Efficiency Estimation from Cobb-Douglas Producion Functions with Composed Error", *International Economic Review, 18*, 435–44.

Meeusen, W. and van den Broeck, J. [1977b], "Technical Efficiency and Dimension of the Firm: Some Results on the Use of Frontier Production Functions", *Empirical Economics, 2*, 109–22.

Meller, P. [1976], "Efficiency Frontiers for Industrial Establishments of Different Sizes", *Explorations in Economic Research*, Occasional Papers of the National Bureau of Economic Research, 3, 379–407.

Mitchell, W.C. [1937], "The Social Sciences and National planning", in *Planned Society: Yesterday, Today, Tomorrow*, ed. E. MacKenzie. New York: Prentice Hall.

Moene, K.O. [1984], "Investment and Fluctuations. Optimal 'Putty-Clay' Investments under Uncertain Business Prospects", Memorandum from the Department of Economics, University of Oslo, No. 9.

Moene, K.O. [1985], "Fluctuations and factor proportions: Putty-clay investments under uncertainty", in *Production, Multi-Sectoral Growth and Planning*, ed. F.R. Førsund, M. Hoel and S. Longva. Amsterdam: North-Holland, 87–108.

Muysken, J. [1979], "Aggregation of Putty-Clay Production Functions", Rijksuniversiteit te Groningen, doctoral dissertation.

Muysken, J. [1983], "Transformed Beta-Capacity Distributions of Production Units", *Economics Letters, 11*, 217–21.

Muysken, J. [1985], "Estimation of the capacity distribution of an industry: The Swedish dairy industry 1964–1973", in *Production, Multi-sectoral Growth and Planning*, ed. F.R. Førsund, M. Hoel and S. Longva. Amsterdam: North-Holland, 43–63.

Nelson, R.R. and S.G. Winter [1978], "Forces Generating and Limiting Concentration under Schumpeterian Competition", *Bell Journal of Economics, 9*, 524–48.

Nerlove, M. [1963], "Returns to Scale in Electricity Supply", in *Measurement in Economics: Studies in Mathematical Economics and Econometrics in Memory of Yehuda Grünfeld*, ed. C. Christ et al. Stanford: Stanford University Press.

Nickell, S. [1974], "On the Role of Expectations in the Pure Theory of Investment", *Review of Economic Studies, 41*, 1–19.

Nickell, S.J. [1978], *The Investment Decisions of Firms*. Cambridge: Cambridge University Press.

Nishimizu, M. and J.M. Page, Jr. [1982], "Total Factor Productivity Growth, Technological Progress and Technical Efficiency Change: Di-

mensions of Productivity Change in Yugoslavia, 1965–78", *Economic Journal, 92*, 920–36.

Norman, G. [1979], "Economies of Scale in the Cement Industry", *Journal of Industrial Economics, 27*, 317–33.

Olson, J.A., P. Schmidt and D.M. Waldman [1980], "A Monte Carlo Study of Estimators of Stochastic Frontier Production Functions", *Journal of Econometrics, 13*, 67–82.

Peck, S.C. [1974], "Alternative Investment Models for Firms in the Electric Utilities Industry", *Bell Journal of Economics and Management Science, 5*, 420–58.

Phelps, E.S. [1963], "Substitution, Fixed Proportions, Growth and Distribution", *International Economic Review, 4*, 265–88.

Pitt, M.M. and L.-F. Lee [1981], "The Measurement and Sources of Technical Inefficiency in the Indonesian Weaving Industry", *Journal of Development Economics, 9*, 43–64.

Prais, S.J. [1974], "A New Look at the Growth of Industrial Concentration", *Oxford Economic Papers, 26*, 273–88.

Pratten, C.F. [1971], "Economies of Scale in Manufacturing Industries", Department of Applied Economics Occasional Papers, No. 28, Cambridge: Cambridge University Press.

Quandt, R.E. [1966], "On the Size Distribution of Firms", *American Economic Review, 56*, 416–32.

Ribrant, G. [1970], "Stordriftsfördelar inom industriproduktionen", SOU 1970:30, Stockholm.

Richardson, G.B. [1960], *Information and Investment*. Oxford: Oxford University Press.

Richmond, J. [1974], "Estimating the Efficiency of Production", *International Economic Review, 15*, 515–21.

Ringstad, V. [1967], "Econometric Analysis Based on a Production Function with Neutrally Variable Scale-Elasticity", *Swedish Journal of Economics, 69*, 115–33.

Ringstad, V. [1974], "Some Empirical Evidence on the Decreasing Scale Elasticity", *Econometrica, 42*, 87–101.

Ringstad, V. [1971], "Estimating Production Functions and Technical Change from Micro Data", Central Bureau of Statistics, Oslo.

Rothschild, M. and J. Stiglitz [1970], "Increasing Risk: A Definition", *Journal of Economic Theory, 2*, 66–84.

Russell, R.R. [1985a], "Measures of Technical Efficiency", *Journal of Economic Theory, 35*, 109–26.

Russell, R.R. [1985b], "On the Continuity of Measures of Technical Efficiency", unpublished (New York Univerity).

Salter, W.E.G. [1960], *Productivity and Technical Change*. Cambridge: Cambridge University Press.

Sato, R. [1970], "The Estimation of Biased Technical Progress and the Production Function", *International Economic Review, 11*, 179–208.

Sato, R. [1975], *Production Functions and Aggregation*. Amsterdam: North-Holland.

Scherer, F.M. [1974], "The Determinants of Multiplant Operations in Six Nations and Twelve Industries", *Kyklos, 27*, 124–39.

Scherer, F.M., A. Beckenstein, E. Kaufer and R.D. Murphy [1975], *The Economics of Multi-Plant Operation, An International Comparison Study*. Cambridge: Harvard University Press.

Schmidt, P. [1976], "On the Statistical Estimation of Parametric Frontier Production Functions", *Review of Economics and Statistics, 58*, 238–9.

Schmidt, P. [1978], "On the Statistical Estimation of Parametric Frontier Production Functions: Rejoinder", *Review of Economics and Statistics, 60*, 481–2.

Schmidt, P. [1985–86], "Frontier Production Functions", *Econometric Reviews, 4*, 289–328.

Schmidt, P. and T.-F. Lin [1984], "Simple Tests of Alternative Specifications in Stochastic Frontier Models", *Journal of Econometrics, 24*, 349–61.

Schmidt, P. and C.A.K. Lovell [1979], "Estimating Technical and Allocative Inefficiency Relative to Stochastic Production and Cost Functions", *Journal of Econometrics, 9*, 343–66.

Schmidt, P. and C.A.K. Lovell [1980], "Estimating Stochastic Production and Cost Frontiers when Technical and Allocative Inefficiency are Correlated", *Journal of Econometrics, 13*, 83–100.

Schmidt, P. and R. Sickles [1984], "Production Frontiers and Panel Data", *Journal of Business and Economic Statistics, 2*, 362–74.

Schumpeter, J.A. [1942], *Capitalism, Socialism and Democracy*. New York: Harper.

Seierstad, A. [1985], "Properties of production and profit functions arising from the aggregation of a capacity distribution of micro units", in *Production, Multi-sectoral Growth and Planning*, ed. F. Førsund, M. Hoel and S. Longva. Amsterdam: North-Holland.

Seip, D. [1974], "A Geometrical Approach to Aggregation from Micro to Macro in Putty-Clay Aggregated Production Functions", Memorandum from the Institute of Economics, University of Oslo.

Seitz, W.D. [1970], "The Measurement of Efficiency Relative to a Frontier Production Function", *American Journal of Agricultural Economics, 52*, 505–11.

Seitz, W.D. [1971], "Productive Efficiency in the Steam-Generating Industry", *Journal of Political Economy, 79*, 878–86.

Shephard, R.W. [1953], *Cost and Production Functions*. Princeton: Princeton University Press.

Simon, H.A. [1979], "On Parsimonious Explanations of Production Relations", *Scandinavian Journal of Economics, 81*, 459–74.

Simon, H.A. and C.P. Bonini [1958], "The Size Distribution of Business Firms", *American Economic Review, 48*, 607–17.

Singh, A. and G. Whittington [1975], "The Size and Growth of Firms", *Review of Economic Studies, 42*, 15–26.

Söderström, H.T. [1976], "Production and Investment under Cost of Adjustment, A Survey", *Zeitschrift für Nationalökonomie, 36*, 369–88.

Solow, R.M. [1962a], "Substitution and Fixed Proportions in the Theory of Capital", *Review of Economic Studies, 29*, 207–29.

Solow, R.M. [1962b], "Technical Progress, Capital Formation and Economic Growth", *American Economic Review (P & P), 52*, 76–86.

Solow, R.M. [1970], *Growth Theory. An Exposition*. Oxford: Clarendon.

Srinivasan, G. and M.R. Fry [1981], "Energy Savings in the Cement Industry", *Process Economics International, 11*, 30–32.

Steindl, J. [1965], *Random Processes and Growth of Firms*. London: C. Griffing.

Steindl, J. [1968], "Size Distributions in Economics", *International Encyclopedia of the Social Sciences, 14*. New York: MacMillan, 295–300.

Sterner, T. [1985], "Energy use in Mexican Industry", Department of Economics, University of Gothenburg.

Stevenson, R.E. [1980a], "Measuring Technological Bias", *American Economic Review, 70*, 162–73.

Stevenson, R.E. [1980b], "Likelihood Functions for Generalized Stochastic Frontier Estimation", *Journal of Econometrics, 13*, 57–66.

Stigler, G. [1939], "Production and Distribution in the Short Run", *Journal of Political Economy, 47*, 305–27.

Summa, T. [1986], "Intra-Industrial Technical Progress and Structural Change", ETLA, Helsinki.

Svennilson, I. [1944], "Industriarbetets växande avkastning i belysning av svenska erfarenheter", *Studier i ekonomi och historia tillägnade Eli F. Heckscher 24-11-1944*, Stockholm.

Svennilson, I. [1945], "Strukturrationalisering", *Harald Nordenson 60 år, En samling uppsatser tillägnade Harald Nordenson*, Stockholm.

Taylor, T.G. and J.S. Shonkwiler [1986], "Alternative Stochastic Specifications of the Frontier Production Function in the Analysis of Agricultural Credit Programs and Technical Efficiency", *Journal of Development Economics, 21*, 149–60.

Timmer, C.P. [1971], "Using a Probabilistic Frontier Function to Measure Technical Efficiency", *Journal of Political Economy, 79*, 776–94.

Todd, D. [1971], "The Relative Efficiency of Small and Large Firms", Report No. 18, Committee of Inquiry on Small Firms. London: H.M.S.O.

Todd, D. [1985], "Productive Performance in West German Manufacturing Industry 1970–80: A Farrell Frontier Characterisation", *Journal of Industrial Economics, 33*, 295–308.

Tyler, G.T. and L.-F. Lee [1979], "On Estimating Stochastic Frontier Production Functions and Average Efficiency: An Empirical Analysis with Colombian Micro Data", *Review of Economics and Statistics, 61*, 436–8.

Vining, Jr., D.R. [1976a], "Autocorrelated Growth Rates and the Pareto Law: A Further Analysis", *Journal of Political Economy, 84*, 369–80.

Vining, Jr., D.R. [1976b], "Capacity Expansion and its Implications for the Size Distribution of Firms: Some Remarks on a Recent Paper by L. Hjalmarsson", *European Economic Review, 7*, 282–6.

Wedervang, F. [1964], *Development of a Population of Industrial Firms*. Bergen: Scandinavian University Books.

Wibe, S. [1980], "Teknik och aggregering i produktionsteorin. Svensk järnhantering 1950–1975; En branschanalys", Umeå Economic Studies, 63.

Williamson, O.E. [1968], "Economies as an Antitrust Defense: The Welfare Trade-Offs", *American Economic Review, 58*, 18–36.

Wohlin, L. [1970], *Skogsindustrins strukturomvandling och expansionsmöjlighet*. Stockholm: Industriens Utredningsinstitut.

Zellner, A. and N.S. Revankar [1969], "Generalized Production Functions", *Review of Economic Studies, 36*, 241–50.

Zellner, A., J. Kmenta and J. Drèze [1966], "Specification and Estimation of Cobb-Douglas Production Function Models", *Econometrica, 34*, 784–95.

Zieschang, K.D. [1983], "A Note on the Decomposition of Cost Efficiency into Technical and Allocative Components", *Journal of Econometrics, 23*, 401–5.

Åkerman, G. [1931], "Den industriella rationaliseringen och dess verkningar", SOU 1931:42, Stockholm.

Subject Index

Author Index

ANALYSES
OF
INDUSTRIAL
STRUCTURE

A PUTTY-CLAY APPROACH

The measurement of industrial structure and productivity growth has a long history in economics. Its usefulness has been limited, however, by the rather simple methods that have been applied.

This book introduces a putty-clay production function in the analysis of technical change. A coherent framework for studying industrial structure is developed.

The optimal production structure is conceptualized within a vintage framework. Frontier and short-run industry production functions are used to measure productivity and technical progress. The usefulness of the method is demonstrated in case studies of various industries.

Distribution:
Almqvist & Wiksell International
Stockholm, Sweden