TECHNOLOGICAL PROGRESS IN ENTERPRISES AND DIFFUSION OF INNOVATIONS

Theoretical reflections and empirical evidence

PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Rijksuniversiteit Limburg te Maastricht,
op gezag van de Rector Magnificus, Prof. mr. M.J. Cohen,
volgens het besluit van het College van Dekanen,
in het openbaar te verdedigen
op donderdag, 8 april 1993 om 14.00 uur

door

Paul Joseph Marie Diederen

UPM
UNIVERSITAIRE PERS MAASTRICHT

Promotores:

Prof.dr. J. Muysken
Prof.dr. F.C. Palm

Beoordelingscommissie:

Prof.dr. L.L.G. Soete (voorzitter)
Prof.dr. P.A. David (Stanford University, California)
Prof.dr. A. van Witteloostuijn

CIP-DATA KONINKLIJKE BIBLIOTHEEK, DEN HAAG

Diederen, Paul Joseph Marie

Technological progress in enterprises and diffusion of innovations:
theoretical reflections and empirical evidence /
Paul Joseph Marie Diederen. - Maastricht:
Universitaire Pers Maastricht
Thesis Maastricht. - With ref. - With summary in Dutch.
ISBN 90-5278-067-6
NUGI 681
Subject headings: innovation diffusion / technological change.

*When the great Rabbi Israel
Baal Shem-Tov saw misfortune
threatening the Jews it was
his custom to go into a certain
part of the forest to meditate.
There he would light a fire,
say a special prayer, and the
miracle would be accomplished
and the misfortune averted.*

*Later, when his disciple, the
celebrated Magid of Mezritch,
had occasion, for the same
reason, to intercede with heaven,
he would go to the same
place in the forest and say:
"Master of the Universe, listen!
I do not know how to light the fire,
but I am still able to say the prayer."
and again the miracle would
be accomplished.*

*Still later, Rabbi Moshe-Leib
of Sasov, in order to save his
people once more, would go into
the forest and say: "I do not know
how to light the fire, I do not
know the prayer, but I know the
place and this must be sufficient."
It was sufficient and the
miracle was accomplished.*

*Then it fell to Rabbi Israel
of Rizhyn to overcome misfortune.
Sitting in his armchair, his head
in his hands, he spoke to God:
"I am unable to light the fire
and I do not know the prayer;
I cannot even find the place
in the forest. All I can do
is to tell the story, and
this must be sufficient."
And it was sufficient.*

*God made man because
He loves stories.*

Elie Wiesel
The Gates of the Forest, 1966

Aan mijn ouders

Contents

Preface

In the Netherlands, the art of painting has a long tradition that has given us Rembrandt and Van Gogh. From this rich tradition, the school of "De Stijl", with painters like Piet Mondriaan, emerged in our century. The paintings of artists of De Stijl are an attempt to condense visual experience, to reduce perception it to its essential elements, straight lines and primary colours. Is this way of looking at the world around us inspired by the world as it has become in our age? Or have we shaped our environment like this, because we started to look at it this way? Whichever is the case, once familiar with the style elements of De Stijl, one tends to recognize them frequently in everyday experience.

Like artists, economists in their models attempt to condense economic reality to what they see as its elementary features. Models impose a structure upon reality and filter out the elements considered inessential. As models find their way into our minds, they determine what we see. We are like painters that train to see three dimensions in terms of two and surfaces in terms of lines. As Keynes remarked, the observers who believe themselves to be most impartial are usually the slaves of some obscure economist.

This thesis is about my attempts to construct a model of economic behaviour, my search for the essentials. It is about what to put in and what to leave out. It was my aim to come up with a Mondriaan painting: three straight lines well placed and two bright colours, clear, unambiguous, austere and brilliant, a picture so plain and simple that when you pass it by in a museum, you look and think: I could have done that (but I didn't).

This book was started more or less by accident and it largely happened to develop in reverse: from the empirics to the theory, from data to concepts, from concrete to abstract, from the complexity of reality to the cleanliness of mathematics. In 1986, after completing my studies at Amsterdam University, I came to Maastricht to participate in a research project with Franz Palm and Joan Muysken on employment effects of technological change. This project got me on the track of diffusion models which ultimately resulted in the last chapter of this thesis. A next project concerned a similar issue in the banking industry, giving me the basics for chapters four and five. As I tried to work out the issue of diffusion, I became interested in the microeconomic representation of technological change, elaborated in chapter three. This book is thus the inverted report of an erratic but instructive journey.

Many people contributed to the genesis of this dissertation. I am very much indebted to my supervisors Joan Muysken and Franz Palm, for stimulating me in making a book out of my capricious ventures, for reading the countless drafts and notes that I produced on my way to the present version, and for helping me with numerous suggestions. I am grateful to the members of the evaluation committee, Luc Soete, Paul David and Arjen van Witteloostuijn, for their helpful comments. I am indebted to Luc also for giving me the opportunity to work within the stimulating environment of MERIT.

My sincere gratitude goes to René Kemp and Rombout de Wit, with whom I closely cooperated on the projects that lay at the roots of this dissertation. I wish to thank Eugène Canjels for his valuable research assistance. I would like to thank Bart Verspagen, Huub Meijers, Adriaan van Zon and all other members of the MERIT 'productivity group' for stimulating discussions and for their comments on draft chapters. My thanks go furthermore to Geert Woltjer, Arie van Lier and Rezaul Kabir for many helpful discussions and for their moral support. I thank the administrative staff at MERIT and at Quantitative Economics for helping me with many day-to-day things that served to make my life as a university staff member a pleasant experience.

I would like to thank Malgorzata Wozniak, who firmly supported me in this project as in many others, and my parents, who first enabled me and later always encouraged me to pursue and persevere.

Maastricht, January 1993

1. General introduction

1.1 Introduction

One of the permanent puzzles of human existence is the problem of free choice: do we really choose when we think we do and to what extent can we determine the course of our life. Is the essence of our existence in our power to decide, to act and to be responsible, or in our aptitude to experience, to undergo and to accept. These questions present themselves in different societies and eras in different guises. They find their way into religion and metaphysics (freedom or determinism), ethics (the meaning of responsibility), studies of history (e.g. the importance of the individual in the historical process), psychology (the role of the unconscious; behaviourism) and other social sciences. Answers to these questions seem to depend on cultural background and historical experience. One is likely to encounter quite different attitudes concerning these matters in India, Poland, the United States or The Netherlands. Questions of freedom of choice are also at the roots of economics: to what extent can the economic process be guided, can the coordination of production and distribution be designed; to what extent can a society choose its economic policies and implement them effectively? Does not the economic system seem autonomous, and therefore maybe deterministic, *because* it is founded on free choice of the individual? Present day economists, more than economists two or three decades ago, tend to be rather optimistic about individual choice, but pessimistic about collective choice. There is confidence in individual rationality, but a lack of confidence in the ability of agents to take efficient decisions collectively and to implement them consequently. The grown confidence in the efficiency of market coordination, the acceptance of theories of rational expectations, the recognition of prisoners dilemmas and the demise of Keynesian macroeconomic policies can be taken as evidence of this.

The problem of the existence of real choice manifests itself in an especially explicit way in economics in the study of technological change. Technology determines how we can transform our material and immaterial environment to our benefit or detriment. It is the long term basis of activity, productivity and wealth. Thus it is important to answer the question whether the speed and course of technological change can be steered and how. Are we the masters or the slaves of technology, does technological progress liberate or subdue us? Scientists, entrepreneurs and managers decide on matters of technology, but they do so given the circumstances which they collectively create. Decisions on technology are often characterized by large external effects, both at present and in the remote future, which are unforeseen at the moment the wheel of invention is put in motion. Once the wheel is running in a certain direction, it seems to get a momentum of its own. This has led many to turn to a form of technological determinism: individually we may choose, but only within the limits set by the way history progresses; technology determines society. But we cannot afford technological determinism, since it may take us down any blind alley. Thus there is every reason to come to grips with the social and economic determinants of the process of technological change, the mechanisms of innovation and diffusion, and to discover in what way they can be influenced.

1.2 Economics and technical change

Economic theory deals with issues of human behaviour, starting from the premises of both unlimited needs and scarcity of resources. Because resources are limited, not all needs can be fulfilled. Therefore agents have to choose, to take decisions regarding how to put available resources to use. A common assumption in economics is that behaviour is rational in some sense and that decision making can be described as optimizing under constraints. Given these assumptions about individual behaviour, an important economic problem is the issue of coordination of the behaviour of individuals in an society. Available resources are in the hands of a variety agents which all simultaneously take decisions on how to put them to use. The decisions of one individual have repercussions on the possibilities of others. Therefore there is in every economy an institutional structure, a mechanism of interaction, by which individual decisions are coordinated. In economics one usually considers coordination of decisions through a market mechanism. Thus the core of economic theory consists of the behavioural assumption of rationality and the institutional assumption of a market coordination mechanism.

Technological change occurs when new and more advantageous possibilities to use the limited resources in society open up. It has a direct impact on the constraints of an agent's decision problem, and therefore on the behaviour that economic theory tries to explain. There are two main questions regarding technological change in this context: first, what economic factors *cause* it to occur and determine its shape, and second, what *effects* does it have on economic behaviour and performance. For a long time, the first question was largely ignored by economists. Technological change was considered exogenous. The second question, however, has gained a place in two parts of economics some decades ago. On the one hand, technological change has become part of production function theory. The level of technology is represented by a parameter in the production function, and changes in this parameter transform the production function over time [see e.g. Coombs, Saviotti and Walsh (1987), Gomulka (1990) or any other textbook on the economic theory of technical change]. On the other hand, technological change is dealt with in a welfare theoretical context. New technologies have to a certain extent a public good character: they can be both non-rivalrous and non-exclusive in their use. Therefore the market may fail to ensure a Pareto efficient level of their provision. Market failure can be seen as a justification for government intervention or regulation [see for analyses along these lines e.g. Gomulka (1990), Stoneman (1987) and Dasgupta and Stoneman (1987)].

Over the last decade, however, some new trends have emerged. Since the beginning of the 1980's, as a consequence of changes in the structure of the world economy and of diverging rates of growth between industries and between countries, a specific range of economic issues has increasingly called for attention. New economic powers seem to emerge and take over leadership in world trade, new industrial and service sectors start to dominate economic development. Growth of industries is uneven and unstable, and fast growing and leading industries often seem to be characterized by large firms, oligopolistic competition and disequilibrium. It proved a challenge to analyse and explain these dynamic features, which could not satisfactorily be met within the framework of traditional analysis. This has led both to efforts to extend the traditional framework into new directions, and to increased critique on the foundations of the whole theoretical structure. Among important extensions are the so called new growth theory and recent developments in industrial economics, in particular the analysis of strategic behaviour, with its applications of game theory. A fundamental critique of mainstream economic theory has been developed within the context of what has become to be called evolutionary economics. It is remarkable that both in the recent extensions of the traditional approach and in the alternatives advanced by economists critical of established views, technological change occupies a more or less central position. There seems to be a general recognition of the importance of the technology factor for the explanation of the above mentioned structural changes and economic dynamics.

Because the modelling approaches in subsequent chapters try to combine elements from both mainstream and evolutionary thought, it is useful to take a closer look at evolutionary economic theory. As mentioned, evolutionary economics has originated from studies of uneven patterns and cycles in economic development. Therefore, the basic questions of this field concern the determinants of economic dynamics. Attempts at an explanation of these uneven patterns of development have led to the articulation of a different notion of the determinants of economic behaviour, a new set of key concepts to analyse economic development. A basic conjecture is that changes in economic activities should be understood in relationship to technological and institutional development. The explanation of patterns of economic growth and structural change should be sought in the development of a technological paradigm, which can be defined as 'a pattern for solution of selected techno-economic problems, based on highly selected principles derived from the natural sciences. A technological paradigm is both a set of *exemplars* - basic artefacts which are to be improved [...], and a set of *heuristics*.' [Dosi (1988)]. A technological paradigm revolves around some key factor [Freeman and Perez (1988)], which is characterized by low and falling costs, nearly unlimited availability, and general applicability in many sectors of activity. One may think of oil in the past decades, or of semiconductors presently. The appearance and development of such a key factor opens up a vast range of profit opportunities through product and process innovations in different industries. It thereby leads to changes in factor demands and demands for skills, to unevenly distributed jumps in potential productivity, to new types of best practise organization of firms, to new forms of (market- or other) coordination of economic activity, to a new structure of the economy and, arguably, even to new socio-political regimes [see e.g. Mathews (1989), Boyer (1988)]. A recurring theme in this literature is the complex relationship between free choice and determinism. Economic development, in this view, is predisposed by its historical context, and current investment and growth opportunities are limited by past development. Economic development is path dependent, a specific path can be chosen by mere fluke and there is no reason to assume that any chosen path is globally optimal.

The dynamic analysis of economic development and technological change is a field pioneered by Schumpeter (1939, 1942) and subsequently developed and extended by a wide range of authors, of which Freeman, Dosi and Nelson and Winter are among the most well known [see e.g. Nelson and Winter (1982) and various contributions in Dosi *et. al.* (1988)]. The approach has been applied to tackle problems and policy issues ranging from microeconomics to macroeconomics and economics of international trade. It is remarkable, however, that evolutionary thought, though rather dissimilar in approach from mainstream approaches in economics, seems close to the theories of some authors starting out from other perspectives. An example would be Scott, who published a theory on economic growth [Scott (1989)].[1] Scott's main contribution is his inquiry into the character of investment, rejecting the definition of investments as physical increments to a capital stock. He alternatively defines investment as "all expenditures undertaken to improve assets (whether human or not), over and above required maintenance", as the cost, in terms of consumption forgone, of changing economic arrangements. By describing investment in this way, all growth results from investment of some kind (abstracting from demographic change), and there is no necessity any more for a separation between investment and technical progress as causes of growth [Scott (1989), section 1.5]. Scott stresses the relationship between investment, investment opportunities and innovation: investment opportunities and opportunities for technological progress are not gradually exhausted, nor are they created by exogenous scientific advance. Rather, "it is undertaking investment which itself creates and reveals the further opportunities" [Scott (1989), section 6.5]. Investment leads to change and change is essential to learning and invention. Application of these views in growth theory leads to a dynamic and history dependent growth model, in which technological progress is an integral part of change.

The evolutionary analysis of economic behaviour also accords to a large extent with views expressed by authors like Porter, writing on strategic management issues [see in particular Porter (1990a&b)]. Porter analyses in his books the process of competition and the determinants of competitiveness of firms. Competition is a dynamic process, a struggle in a constantly changing environment. Competitiveness is the sustained capacity to generate income from producing and marketing products. Like evolutionary economists, Porter puts technological change at the centre of his analysis, stating that 'competitiveness depends on the capacity to innovate and upgrade', and that 'the basis of competition has shifted more and more to the creation and assimilation of knowledge' [Porter (1990b)]. Sustainable rates of innovation, leading to cost reductions through rises in productivity and, more importantly, to product differentiation, are the key to profitability. Porter argues that a sustained flow of innovations needs a conducive and challenging environment to come about and be maintained, and then uses hundreds of pages to analyse in great detail which factors make up the conditions which favour innovative activity. The main message is that competitive pressure is vital to innovative activity, and that therefore government policy should be directed at maintaining a competitive environment.

Over time, the development of this evolutionary analysis of economic cycles has broadened and instigated a more fundamental discussion on the appropriateness of the usual premises of economic theory. To see the relevance of this discussion, it is useful to trace the reasoning that leads to the argument. The starting point is the recognition of the central importance of innovative activity for competition and the course of economic development. Innovative activity is any activity directed at rising the earning capacity of a firm by introducing a novelty of some sort. It requires the utilisation of scarce resources, and therefore falls within the realm the theory of decision making. The question of central concern is thus, to what extent innovative activity can be analysed with the tools of economic theory. The main problem here, as distinguished by evolutionary theorists, is the character of information in innovative activity. There are two aspects: first, the information that has to be dealt with in (deciding on) innovative activity is generally vast and complex, and second, there is genuine uncertainty (in the sense of Knight

[1] Scott's basic model, although developed using other mathematical tools, is quite akin to the model in chapter 3 below, as can be seen by comparing his basic assumptions to the ones advanced here [see Scott (1989), especially chapters 3 to 6, or the summary of results at the beginning of the book].

(1965)) in innovative activity. The latter assertion means that it is impossible to imagine the likely outcomes of innovative activity and the path of future technological development, because essential information is lacking.

Consequently the assumption that decision making on innovative activity can be described with the help of the concept of rationality or optimizing, becomes subject for debate. If in deciding on innovation there is so much information involved that to consider all of it in detail would be prohibitively costly, and if there is vital information lacking which can only be produced by carrying on with investing in innovation, then one may ask in which sense decisions on innovative activity can be optimal. Rather they may be better described as determined by routines and by bounded rationality. This introduces components in decision behaviour which are historically determined, which are by nature unsystematic and unpredictable if considered separate from their historical context.

If agents would be able to optimize, then (disregarding exceptional circumstances) markets will coordinate decisions such that an equilibrium over time will appear: a situation in which no agent is willing to change his course of behaviour, given the circumstances and the behaviour of all other agents. A persistent equilibrium requires that there are no unpredictable future feedbacks from current action. Agents can estimate the possible consequences of their actions and their likelihood. If behaviour is not optimizing, then there is no reason to assume that coordination of decisions through a market mechanism will lead to equilibrium. Actual behaviour is likely to be non-equilibrium behaviour. The coordination of the market will induce agents to adjust their actions, but there is no guarantee of optimality or stability. Thus there is likely to be continuous disequilibrium and adjustment. The market does not coordinate decisions to be optimal, but acts as a selection mechanism for the relatively better decisions.

Finally, if actual decision making is non-equilibrium behaviour, which is by its nature not predictable or systematic, seen separately from its historical context, then it is questionable whether decision making should always be the starting point of economic analysis. If there is a fundamental indeterminacy in decisions that are taken, why then should an economic theory about the behaviour of aggregates allow for a decision theoretic, microeconomic, underpinning. If the purpose of the exercise is not so much to provide an explanation of economic decisions, as to explain economic activities of production, trade and consumption, then to start *not* from a theory on decision making might be more fruitful. Thus it is not clear *a priori* that it is more satisfactory to take firm behaviour as unit of analysis, rather than e.g. a technology.

Leaving this discussion here for what it is, it may be illuminating for the chapters to come, to distil from it two different stylized sets of characteristic assumptions, the first of which could be associated with the mainstream approach and the second with the evolutionary approach to economics (see Table 1). The characterizations are tentative and may be seen as ideal types, with many actual theories and models falling somewhere in between [compare e.g. Silverberg (1988)]. Table 1 may be seen as a framework in which the theories and models in subsequent chapters find their place.

Technological change has become a more important factor in mainstream theories, that is more and more being modelled as endogenous rather than exogenous, especially in long run growth models [see e.g. Lucas (1988), Romer (1990)]. In evolutionary thought its role is different, and it can be said to occupy a central position. Because of that, and because of the complex and uncertain character of new technologies, genuine uncertainty and suboptimal decisions are a vital ingredient of evolutionary thought. Within the framework of mainstream theories, genuine

Table 1: Stylized differences in approaches towards economic theorizing.

	mainstream thought	*evolutionary thought*
technological change	important	endogenous, key variable
behavioural ass.	optimizing, rational	routines, bounded rationality
ass. on information	determinism or risk	complex information, genuine uncertainty
unit of analysis	firm or agent decisions	decisions or other variables
system properties	equilibrium no unpredictable feedbacks	disequilibrium, dynamic adjustment unpredictable feedbacks
role of time	ahistorical	historical, path dependent

uncertainty cannot be analysed, and therefore at most risk (in Knight's sense) is assumed. Suboptimal decision behaviour and lack of information about future likely consequences of present decisions do not yield a market equilibrium, and thus evolutionary analysis concentrates on dynamics, adjustment mechanisms and path dependencies.

1.3 Purpose and outline of the thesis

This thesis is about constructing models of technological change. A model is like a pair of glasses to look at the world, an instrument to understand certain features of economic processes. Constructing such a model is an exercise in analyzing reality, in an attempt to discern its essential characteristics. It is a matter of choosing what to put into the model and what to leave out. Much attention in this thesis is devoted to these choices. It involves first of all clarifying a number of definitional issues: a workable definition of technological change, the nature of terms like equilibrium, optimizing behaviour, dynamics. Secondly, it involves distinguishing key variables that determine the process, like e.g. capacity, inputs, investments, prices. Finally, it concerns making the causal relationships between variables explicit.

The thesis is restricted to certain aspects of technological change. It deals mainly with the diffusion stage of the process of technological change, more specifically, with the diffusion of process innovations. It does not deal explicitly with the generation of innovations, but concentrates on models describing the consecutive adoption of new process technologies by business enterprises. An important feature of this research is that it tries to deal with two related phenomena simultaneously. First of all, there is the familiar observation that innovations diffuse slowly through an industry, that firms adopt the same innovation at different times. Secondly, there is the observation that over time firms adopt different innovations. We shall try to deal with these two features of reality in an integrated framework. Innovations are thus seen, not as isolated phenomena, but as steps on a technological track, on which firms move from different starting positions and with different speeds. On the one hand, there is the movement of innovations through an industry, and on the other hand, there is the movement of firms through a range of innovations.

The main questions addressed in this thesis are about understanding the role of technical change in determining economic behaviour and capturing this in a model. To start with, there are a number of analytical questions on how to understand certain assertions about technological change and translate them into a model. Important assertions concern the endogenous nature of technological development, the path dependent nature of progress, the complexity of information and bounded rationality, and the determinants of the dynamics of adjustment to new technological opportunities. These questions will be dealt with by developing models of technology adoption and diffusion. Related to these issues are questions concerning the consequences of assumptions of endogenous technical progress, path dependence and bounded rationality for growth rates and patterns of industry development. To some degree models will be helpful to evaluate these consequences. Complementary to questions of analytical nature, there are empirical questions about the relative importance of supposedly determining factors for actual processes of technological change. Answers to such questions are sought by analysing a 'case' of technological change, the introduction of a number of innovations in a group of Dutch banks over a period of nine years, and by estimating (some versions of) the theoretical models.

An interesting aspect of these matters, to which empirical research may give an answer, is the question at which level of aggregation economic development is most regular and predictable. In particular, there is the issue of the importance of rationality and predictability at the microlevel for orderliness at the aggregate level: is it helpful to model innovation adoption and then aggregate, to describe diffusion accurately, or can the diffusion curve better be predicted by reference to aggregate variables directly? It could be that, though technology adoption is rational and deterministic at the microlevel, it may lead to irregular patterns at the aggregate level, due to externalities and unpredictable interactions and feedbacks. It could also be that, though innovation adoption may be irregular and unpredictable at the microlevel, it may be regular at the aggregate level, due to mutual feedbacks, the averaging out of noise, adjustment and selection mechanisms. In that case, a certain pattern of technology adoption is followed, but one cannot predict which firm plays what role in this process.

The treatment of the questions above is to be considered within the context which has already been sketched in section 1.2 above and in Table 1. The issues will be approached using both concepts and techniques from mainstream economics, but in doing so an attempt will be made to take advantage of insights from evolutionary economics. The body of the thesis consists out of five chapters. The argument starts out, in Chapter 2, with an attempt to put technological change in a broader perspective. The topics introduced above are elaborated and an overview over issues concerning innovation diffusion and adoption which commanded attention of other researchers in the field is presented, with references to previous literature. A representation of two models, a firm model of induced innovation and the epidemic diffusion model, is given, because they will serve as a starting point for models in later chapters. In Chapters 3 and 6, two main models are developed and explored. These models fall into different categories, as described by Table 1. Chapter 3 is an attempt to use the basis and the techniques of mainstream economics, but to insert into this framework a number of elements drawn from evolutionary thought. In particular, a dynamic model is developed which describes the investment planning of firms, confronted with opportunities to invest in technological progress. The firm is assumed to optimize an objective function under technological and market constraints. This model contains some elements of bounded rationality and routinized behaviour, disequilibrium dynamics, dynamic increasing returns and path dependence.

The model is tested empirically, using a database from a Dutch banking organization. In Chapter 4, recent technological developments in banking are briefly described, and the database used to test the models is introduced. An inductive analysis of automation, size and costs in the case study banking organization is given and evidence of scale economies is explored. Chapter 5 deals with empirical tests of some aspects of the model in Chapter 3. The data available contain only short time series, but there is a lot of cross section material. In Chapter 5 the model is adapted to fit the restrictions posed by the database, and consequently a simplified version of the model is estimated.

In Chapter 6 the modelling problem is approached from a different angle. The models used there take a specific production technique as unit of analysis, there is no reference to equilibrium, and economic behaviour is represented as some type of an adjustment process. These models have some features stressed by evolutionary thought, but they are not evolutionary models in a narrow sense, in that they do not contain an explicit selection mechanism, describing how less efficient firms are forced into bankruptcy, nor a chance mechanism generating innovations at unpredictable moments. The type of model explored there does take regularity at the aggregate level as its point of departure, however, without making explicit assumptions about optimizing behaviour at the micro level. Results of empirical testing of the models of Chapter 6 are reported in the same chapter.

The thesis ends with a summary and some general conclusions. The main theoretical results of the study concern, first of all, the analysis of investment planning by firms that operate in a market where opportunities to invest in productivity rise are important. The relationship between technological opportunities, market constraints, planned growth of output and planned productivity rise are illuminated. Secondly, results concern the process of competition in a market by heterogenous firms that invest in technical improvements. Some light is cast on the relationship between firm size, market structure and the speed of technical progress, and on the issue of steady state growth. The main empirical results pertain to the mechanisms of diffusion of innovations in banking. It is found that adoption behaviour shows regular patterns but is weakly related to cost data.

2. Theoretical background

2.1 Introduction

This chapter will be used to render the perspective which underlies the theory and models in later chapters explicit. The starting point, in section two, is a general sketch of economic behaviour, in relationship to the emergence of technological change in an economic system. The third section deals with the relationship between the adoption of innovations and the process of diffusion, and highlights a number of characteristics of the diffusion of innovations. The fourth section considers two models describing the introduction of innovations. These models originate from two different approaches towards modelling and consider innovation at a different level of aggregation. The fifth section highlights a number of distinctive issues on which different authors dealing with the adoption and diffusion of innovations have taken different stances.

2.2 Technological change and economic development.

For many decades, the explanation of technological change has been markedly absent from mainstream economic analysis. Paradoxically, this was the case in an age in which actual technological progress was far from stagnating, and in which the technology factor became an ever more decisive determinant of economic success. Economic history of the last century is a chronicle of launches of innovative products, revolutions in process technology, new developments in work organization, and intensified search for whatever is new and can be sold. In this very period, technical change moved to the background in economic theory. It seems that technology did not slip out of economic analysis because of a lack of empirical relevance. Rather, it disappeared because it gradually drifted out of the economist's analytic perspective. The development of economic analysis and of its main instruments has taken such a course that technological change has shifted from being a variable at the centre of economic analysis, which it was for Ricardo, Marx and Schumpeter, to being an exogenous parameter, a phenomenon at the margin of the realm of economic analysis.

This marginalization of the technology factor may be partly explained by the prominence over the last decades of the short run unemployment problem on the economist's agenda. Another reason may be the popularity of an ever more mathematical approach towards economic analysis. The tools of mathematics often seem to resist to be used for the analysis of qualitative progress, discontinuities in behaviour and unpredictable changes in e.g. technological opportunities. But focus on problems of unemployment and the momentum of the development of the economist's mathematical tool kit may be only part of the explanation. The marginalization of technological change is probably also related to the fact that the study of allocation through the operation of the market mechanism has become ever more firmly established as the core of economic analysis. This focus on market coordination diverts attention from activities taking place *before* economic agents enter the market, from the stages in which investment and production are planned, expectations are formed, innovations are developed, and production takes place. In these pre-market stages of economic activity, technological change takes shape. Here, on shop floors, in laboratories and in board rooms of business enterprises, other social mechanisms operate than the market mechanism, other types of information are available and applicable to decision making and other constraints are relevant. The marginalization of technology in economic analysis is a consequence of the paradigmatic custom of economics to put market exchange at the centre of analysis. Though fruitful for many purposes, this diverts attention away from some essential aspects of behaviour and coordination, when exploring the relationships between technological developments and long term economic changes.

2.2.1 An economic system

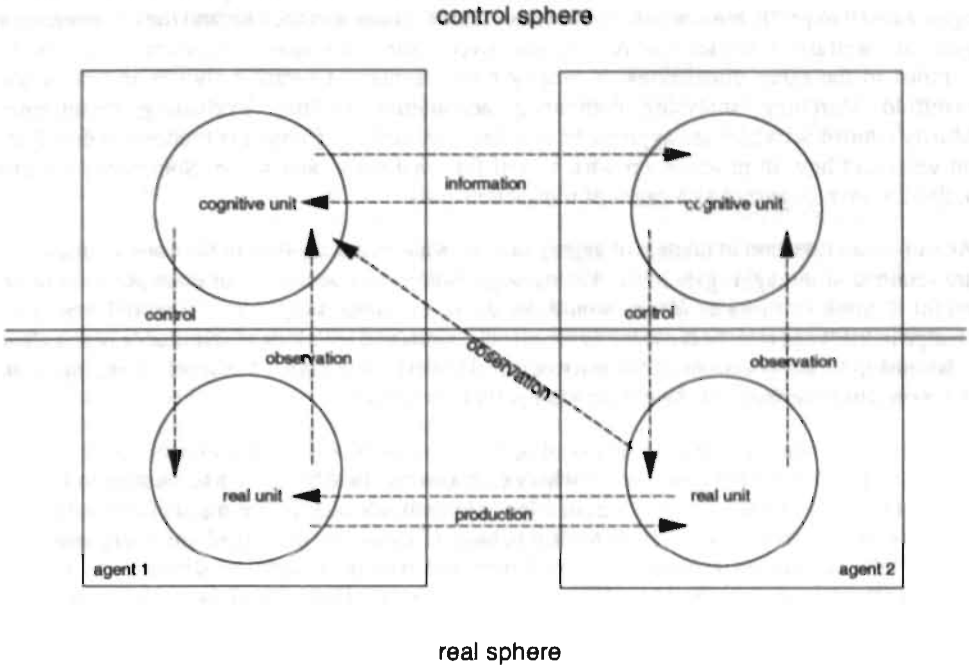
To put the above propositions, it may be helpful to start out from a very general description of activity in an economic system. Any economic system can be thought of as a system in which agents act in relationship to each other. Agents, decision makers at the lowest level of aggregation, are individuals that both think and do. Economists are predominantly interested in what agents do, in terms of production and distribution of the yield, but this is closely related to what and how they think. This can be schematized using Figure 1 [this scheme is used by Kornai (1971), chapters 4 and 5]: in an economic system, every agent is active in two spheres, in the sphere of real activities and in the sphere of cognitive activities. In the real sphere there is production, trade and consumption, in the cognitive sphere experiencing, wanting and decision making. In the cognitive realm, one can distinguish between motivation and control activities. The agent's motivation is his urge to fulfil his needs, which can be of many different types, ranging from food and shelter to appreciation, status and knowledge. The individual tries to attain his goals, satisfaction of his needs or maximization of his utility, by acting in the real sphere: he searches for opportunities, spends time on work, produces, exchanges and consumes. These activities are directed and controlled through cognitive activity: evaluating, checking, deliberating. Real sphere activities require cognitive activities to monitor and control them.

Both in the cognitive and in the real sphere, agents interact. There is exchange of information, through observation and communication, in the cognitive sphere. There is exchange of goods and services in the real sphere, often through institutions like markets. The market mechanism is a particular type of exchange mechanism in the real sphere.

Economic analysis is generally concerned with the explanation of activity in the real part of an economic system: production, exchange of products and consumption. Real activity is visible and can be measured in terms of quantities of product, hours of labour, trade volumes and the like. Activity in the cognitive sphere, on the contrary, can hardly be seen and measured. However, it must be presupposed to understand the activities that take place in the real sphere.

Starting from this representation of an economic system, analysis must proceed in two directions to arrive at an economic model. On the one hand, structure and detail has to be added to the general description of the economic system: one has to assume an institutional setting, property rights, markets, organizational structures like firms, countries, governments, and for some

Figure 1: general representation of an economic system



purposes social classes or auctioneers. On the other hand, one has to introduce useful abstractions and aggregations which do not obscure the essential features of the processes under study. Usually there is a trade off between abstracting of variation in one aspect against abstracting of variation in another, and there is a trade off between detail and tractability of analysis. There is a certain risk to abstract from elements which are decisive for the process one wants to explain. Considering Figure 1, it seems that there are at least five types of abstractions, five dimensions over which aggregation can take place. One can aggregate over agents, production (processes and products), motivation, control processes and time.

Aggregating over *agents* can be done up to different levels: firms, industries, aggregate supply. A firm is an aggregate of agents and, although we often use the assumption that the firm decides and produces, it is in fact a conglomerate of agents that search, ponder, deliberate and argue, and finally decide, produce and sell. Likewise, *products* are commonly aggregated into one homogeneous output and *production processes* are aggregated by the use of one single production function. Another common form of abstraction is aggregation over *time*. One aggregates the time it takes for the economic process to adjust to some change in parameters into one single instant. The issues of aggregating over agents, production and time have defined the border lines between micro and macro theory, and between static and dynamic analysis.

Then, moving to the cognitive sphere, abstraction of variety in *motivation* takes different forms. Different consumer needs can be aggregated into some utility concept, firm goals can be aggregated into profit, but also into market share, sales volume and the like, and the government's goal into welfare or re-election. Aggregating over *control activities* in economic analysis, in contrast to the other dimensions, is usually rather implicit. Control activities in reality are manifold: searching, analysing, computing, negotiating, deciding, evaluating, monitoring. Mostly control activities are aggregated into decision making. Control is reduced to deciding, on what and how to produce, on what to sell for what price, and so on. Sometimes control activities are recognized as a cause of transaction costs.

An important question in matters of aggregation is: what characteristics of the constituting units are retained at the aggregate level. Taking aggregation over agents as an example, it is often useful to think of firms *as if* they would decide, or of labour supply *as if* it would react to a change in the wage rate, because the aggregate decisions and reactions are similar or bear a clear relationship to the decisions at the microlevel.[1] However, the higher the level of aggregation, the more characteristics of the constituting parts may be lost.

Another important question in matters of aggregation is: how does the loss of variety at the aggregate level affect the analysis. To what extent can e.g. firm behaviour be understood, if we abstract from the fact that the firm is a coalition of individuals with different and often conflicting interests. And therefore, what parts of firm behaviour cannot be explained any more when we do not take account of principal-agent relationships within the firm, of divisions of power, bureaucratic procedures, rights, obligations and responsibilities. And at a higher level of aggregation, it can be asked to what extent competition can be understood, if we abstract from variety among suppliers of a market. Can we explain prices if we aggregate supply and do not take account of rivalry.

Issues of aggregating over agents, production, time and firm goals have inspired the development of new approaches, to overcome restrictions caused by certain types of aggregation. Limitations of models where individuals are aggregated into firms have given rise to principal agent theory, and limitations of models of aggregate supply have stimulated the use of non-cooperative game theory in industrial economics. There are models of product diversification, market segmentation, quality competition, alternative firm objectives and so on. A lot of effort goes into microeconomic underpinning of macroeconomics, and into construction of dynamic models.

Aggregating over control activities, however, has long gone without much debate and controversy in the economics discipline. These issues are touched upon in theories of bounded rationality, but have gone without much explicit formalization in the economic literature so far.[2] In the context of an analysis of technical change, it is necessary to reconsider the aggregation

[1] Methodological individualism holds that aggregation over agents distorts our view on economic processes to such extent that our understanding is severely hampered. In Schumpeter's words: "It keeps analysis on the surface of things and prevents it from penetrating into the industrial processes below, which are what really matters. It invites a mechanistic and formal treatment of a few isolated contour lines and attributes to aggregates a life of their own and a causal significance that they do not posses." [Schumpeter, 1939, p.44]

[2] One can refer to Atkinson and Stiglitz (1969) and Nelson and Winter (1982) as examples of models where the concept of bounded rationality is invoked as a cause for the localized character of technical progress.

over control processes, because the variety and complexity of these cognitive activities seem to be decisive factors in the process of technological development. Technological development originates in the control sphere of the economic system, results from cognitive activities.[3]

The most far reaching aggregation of control activities is to aggregate all control processes into decision making, and at the same time to aggregate all information relevant for decision making into prices. These two abstractions are related: if all relevant information is contained in prices, then problems of control easily reduce to decision making. The model of the economic system simplifies in three important respects. First of all, prices are public knowledge. Thus we abstract from information which is not accessible to every agent. Equilibrium prices then truly reflect the scarcity value of tradables, and that is all that is relevant to know to take a decision on production or exchange. Secondly, prices are quantitative measures. There is no problem of relating different types of information to each other, because all information is measured on the same scale. Prices can be directly compared. Thirdly, information in the form of prices involves no complexity and can be readily evaluated. There is no training, skill or research involved in understanding all information. If information is only price information, control comes down to arithmetics, and is therefore relatively easy and cheap.

Abstraction in the realm of control processes may obscure some important features of technical change. The development and use of technology is to a large extent a control process: it is getting to know how to make or do something. This type of control activity is different from decision making by calculating. It is a process that works with technical information, beside price information, which has rather different characteristics. First of all, technical information is not readily accessible. Secondly, it is complex and cannot be reduced to a single dimension. Thirdly, it requires research and training to be able to put technical information to use. Thus control processes, when technological information is involved, are not limited to computing and deciding. There is also gathering and researching of information, comparing and evaluating, studying and mastering, training and developing skills and building up routines. These are all processes that require an investment of time and effort. Thus cognitive activities acquire the characteristics of a production process: inputs are transformed into outputs, which are, beside decisions, knowledge and skills, inventions and innovations.

In this context, it is worthwhile to refer to an alternative to the usual rationality concept, the concept of bounded rationality. Simon, who first introduced this term, argues that, due to cognitive constraints, rationality in economics is restricted: "In any realistic description of the environment of a human decision maker, the variables and information to which he might attend (and to which he must attend to satisfy the strict requirements of rationality) are innumerable. The hypothesis of bounded rationality claims that human beings handle this difficulty by attending to only a small part of the complexity about them. They make a highly simplified model of the world, and they make their decisions in terms of that model and the subset of variables that enter into it. Now this approach may work very well (...) if the number of very important variables is small at any given time, and if this list of important variables does not change from time to time without the change being noticed." [Simon (1986) p. 33-34]. Simon

[3] Mueller argues that our understanding of the functioning of the business corporation has suffered, not so much from lack of recognition of constraints, but from the unwillingness of economists to adopt a more realistic set of behavioural assumptions about managerial motivation. He supports his attack on the assumptions of profit and shareholder wealth maximization and his plea for broader models of managerial motivation with empirical evidence from principal/agent and industrial organization literature [Mueller (1992)].

thus points out that, in taking decisions, human beings use some type of model of the world which is a simplification in two respects. In the model, not only the relationships between the variables in the model are a simplification of reality, in the sense that there is a limited representation causal links, but more importantly, there has been an a priori selection of the variables which enter the model. A lot of information has been left out of consideration, on the a priori assumption that the costs of including this information in decision making, and of broadening the range of possible outcomes, would complicate the process of taking a decision to such an extent that the extra costs would outweigh the benefits. This procedure works well and is efficient, so long as the model is not a poor approximation and as long as the structure of the decision problem does not change, in the sense that formerly unessential variables which are external to the decision model become of great importance, while the decision making procedure is not adjusted.

Bounded rationality points at the limitations of human cognitive capacities. In taking decisions on the introduction of new technology, these limitations are of influence, because these decisions are usually complex in a number of respects, and different types of variables can be relevant. There is usually a lot of technical information to be dealt with which can be complex and veiled in uncertainty. Moreover, the adoption of new technology is often tied up with strategic issues and decisions on the long term course of the company.[4] If bounded rationality is a concept, relevant to the description behaviour of firms with respect to their day to day business, then it is certainly relevant to a description of firm's behaviour with respect to new technology.

This brings us back to the two important questions about matters of abstracting or aggregating: what characteristics of the constituting units are retained at the aggregate level, and how does the loss of variety at the aggregate level affect the analysis. First of all, it is clear that if one considers decision making on the basis of price information as the only type of control process in the economic system, a lot of costly and time consuming activity in the economy moves out of sight. Since comparing readily accessible price information and determining the best offer is relatively easy and cheap, in comparison to searching and evaluating, there would be no good reason why control processes should take lots of time and be subject to capacity constraints. It would be difficult to understand, why all decisions are not taken instantaneously and why routines develop. However, in reality a lot of cognitive activities do not take the form of computing, but of searching and learning. This requires time, capacity and investment of scarce resources. The production of new technology is an example of a cognitive activity that requires investments in searching and training. Capacity constraints in this field of control processes seem to determine the speed and scope of progress. Secondly, variety is an important element in the competitive struggle: firms try to outperform competitors by developing positive differences. Variety in technological capacity, in control and search routines, in command over information and in ability to create is an important element in the determination of the dynamics of competition. Abstracting from the variety in constraints in cognitive capacities takes away an important explanation of economic development.

[4] Freeman (1982, p.149) distinguishes between technical uncertainty, market uncertainty and general business uncertainty.

2.2.2 Summary and implications for modelling

Comparing the real sphere and the cognitive sphere of the economic system, we see there is similarity and difference. A similarity, which is often disregarded, is the fact that in both spheres production takes place and is constrained by capacity limits. In the cognitive sphere, beside decisions, information, knowledge and technology are produced, using human energy and time. An important difference between the two spheres, also often disregarded, lies in the fact that exchange in the real sphere concerns appropriable commodities, and transaction costs are mostly relatively minor, whereas exchange in the control sphere concerns information, which is often less appropriable, and which can involve high transaction costs. To acquire information *per se* can be costless, but to acquire the skills and knowledge necessary to absorb information can require large investments in training and study.

There is a tendency in economic analysis to abstract from the complexity and variety of cognitive activities in the economic system, and, in parallel to this, to aggregate all or most information into prices. On the one hand, this is the consequence of a concentration on real sphere economic activities. On the other hand, this is caused by a focus on issues of allocation and the operation of the market mechanism. In this way activities leading to technical change have tended to get marginalized in economic analysis. The reduction of cognitive processes to choice processes leads to a neglect of important constraints in economic development. It diverts attention from investments needed to pursue these cognitive activities, investments in making choices, in producing and searching for new information, and in learning to reap the benefits of information. These investments are both large in value terms and extensive in terms of time. Moreover, to overlook the importance of these complex cognitive processes leads to a blurring of our perspective on factors influencing competition.

On the topic of constructing models of economic development, in which changing technology is important, some conclusions can be drawn from the exposition above:

1. In an economic system, agents not only produce goods and services (or equipment to produce goods and services), but also information, capacity to deal with information and capacity to produce information. Information, knowledge and technology are output from a largely cognitive production process; agents invest in study to master knowledge and they invest in their ability to develop new knowledge and technology. So there are two types of production processes in the economy, production of goods and services and production of knowledge and technology. In models of economic development over longer periods of time, it is important not to abstract from the last.
2. Cognitive processes, production and processing of information, are costly. Production of technical knowledge and skills requires investment of scarce resources.
3. Time is an important factor. Capacity to produce technical knowledge and skills, to generate information and to reduce technical and economic uncertainty is restricted by human capabilities to learn, to invent and to develop.
4. Although the production *process* in the cognitive sphere of the economy is similar to that in the real sphere, in the sense that both require inputs to yield outputs, the *product*, new information, is of a different character from the products in the real sphere of the economy. An important difference is the lesser degree of appropriability of information which, however, often goes together with relatively high transaction costs. These features complicate market exchange of cognitive output.

5. Production of goods is preceded by and dependent upon production of knowledge and technology. A change in technical knowledge can lead to a change in the capacity to produce goods.

This completes the sketch of the general framework. These conclusions may later be referred to when considering the models of chapter 3. We now turn to a descriptive analysis of the introduction of new technology in enterprises.

2.3 Adoption and diffusion of innovations

Economic systems go through changes over time. There are institutional changes, changes in the volume and composition of production, and consequently in prices, in the allocation of resources and in the distribution of returns. One of the important causes of change is technological progress: changes in the process of transformation of inputs into output. This section deals with an outline of this process of technological change.

2.3.1 Aspects of technical change

It is traditional to distinguish between invention, innovation and diffusion as the constituting parts of the process of technological change. Innovation is defined as the first commercial application of an invention, and diffusion as the spreading of the innovation. This distinction suggests that there is an innovator that bears major costs of research and development, launches a new product on the market and tries to recoup his investment before others enter. Potential followers see that a new market has opened up and try to enter at low costs, taking advantage of the research and development expenses of the innovator. Though conceptually straightforward, this partition of the industry between one innovator and a herd of imitators is of limited practical significance. It suggests a large difference between the first firm to exploit an innovation and later firms. If the innovator is just one out of a number of firms that follow a similar offensive strategy, then the difference between the innovator and early followers will be small, often smaller than between early and late followers. The more proprietary a new technology is, or the more effectively it is kept secret, the less followers can take advantage of R&D efforts of innovators. However, only in few instances an innovator's lead is protected effectively by a patent [Nelson (1988)]. Commonly an innovators lead is challenged quickly by competitors that follow a similar first mover strategy. Early followers often bear costs of R&D which are comparable to those of innovators, run similar risks and also enjoy above normal profits, once they manage to enter the market. Also, early adopters often contribute substantially to the development of a 'dominant design' and to further diversification [see e.g. Abernathy (1978)].

In the case of many process innovations, the origin of the innovation is not one of the competitors in a market, but a supplier of capital goods.[5] In such cases diffusion is promoted by the capital supplier and there is even less room for differences between the first mover and later adopters. Thus, there is no essential difference between the characteristics of innovators and of followers as such. Rather is there a continuum of firms pursuing different strategies and setting different targets [see e.g. Twiss (1986), Martin (1984)]. The spectrum ranges from firms that are inclined

[5] Product-related R&D is estimated to account for 75 to 90% of total R&D expenditures in manufacturing [Dosi (1991, p.188).

to venture early innovation to ones that tend to imitate, differentiate or popularize a new product. Generally, earlier firms run higher risks, invest larger amounts and have higher chances on higher returns than later firms. However, competition makes that not all can reach their targets simultaneously. In principle all firms invest to create progress, but some invest more and more effectively than others, and as a consequence have an earlier result and a higher return.

Another traditional practice is to distinguish between adoption and diffusion. Both terms refer to the introduction of new technology, but each from a different perspective. Adoption refers to the introduction of an innovation as a matter of choice, from the perspective of an agent or firm. "Adoption analysis considers the decisions taken by agents, typically organizations such as firms, to incorporate a new technology into their activities. It is concerned with the process of decision making, and leads to propositions linking the nature and timing of adoption decisions to specified characteristics of adopters, e.g. the size of firms, or their sociometric position within a communications network." [Metcalfe (1988, p.561)]. Adoption is a choice referring to a moment in time. Diffusion, by contrast, refers to the introduction of new technology in firms from an aggregate perspective. Diffusion analysis is not directly concerned with the explanation of decision making, with behaviour of an economic agent, but: "Diffusion analysis is concerned with how the economic significance of a new technique changes over time." [Metcalfe (ibid.)]. The analysis centres, not on the agent, but on the new technology. According to Rogers, "Diffusion is the process by which an innovation is communicated through certain channels over time among the members of a social system." [Rogers (1983, p.5)]. Thus the concept of diffusion does not refer to a choice in time, but to a process over time.

Obviously, adoption decisions and diffusion phenomena are related. By aggregating adoption decisions at every moment in time, one arrives at a diffusion path. A question which is important for modelling is now, whether diffusion is to be conceived of as a coherent process. To understand a sequence of events as constituting a (historical) process implies that one assumes that there is some direct or indirect causal relationship between the events that follow upon each other in time. Thus the question is, whether and to what extent successive adoptions of new technology are causally related to each other, to what extent agents take account of the adoption decisions of others when taking a decisions on technology adoption themselves. It may be that a smooth diffusion curve results from adoption decisions that are not directly causally related: there is no feedback from an adoption decision to the conditions for adoption of other potential users. It may also be that adoptions are directly related: the driver of diffusion is endogenous to the process itself. There could be a feedback through a spread of information, leading to a bandwagon effect, through network externalities, or through markets for inputs and outputs, which would affect the profitability of the innovation relative to the old technique. This determines whether it is possible to explain and model diffusion as a function of aggregate variables, or whether it can only be explained by explicit aggregation of underlying behaviour. It determines whether there is some internal logic or coherence to the process of diffusion, or whether the diffusion path is just the sum of the constituting parts. If adoption decisions are not interrelated, then diffusion patterns are in a sense accidental (exogenous) in shape and can only be modelled by aggregating models of individual firm decisions. If individual adoption decisions are strongly related, if every adoption decision is largely determined by foregoing adoption decisions, then the diffusion process can be described by reference to its own history. This would allow for relatively simple models, avoiding the problem of explicit aggregation.

In the first case, modelling and explanation can follow another course than in the second case. Using the terms of Mohr, there are two types of theory, two modes of explanation, variance theory and process theory [Mohr (1982)]. Variance theory is the archetypal theory in science. "In variance theory, the precursor (X) is a necessary and sufficient condition for the outcome (Y)." In other words, the variation in the dependent variable (Y) is explained by the variation in the independent variables (X), in the sense that variation in variables (X) are both sufficient and necessary for variation in (Y) to occur. If adoptions are independent of each other, variation in adoption decisions (Y) at any moment in time must be explained by variation in firm characteristics (X) at some moment in time.

According to Mohr, there is another type of theory, called process theory, that can also be accepted as a type of explanatory theory. A process theory states necessary conditions, in this case the existence of potential adopters and an innovation, but does not explain the adoption of this particular innovation by that particular firm by recourse to a sufficient cause. "In process theory, the precursor (X) is a necessary condition for the outcome (Y)." The precursor is not a sufficient condition. If there is no innovation, then there is no adoption, but not: if there is an innovation, then it is adopted by this or that potential user. However, since we are after an explanation of (Y), we are concerned with when (Y) occurs, not when it does not. The fact that the precursor is necessary for the outcome is insufficient to constitute an explanatory theory. In process theory, identification of necessary conditions is supplemented by a probabilistic process that specifies the chance that the one necessary element, the innovation, is linked to the other, an element out of the population of potential adopters. Together, the necessary conditions plus the chance process explain that in time firms adopt the innovation.

Mohr expresses the difference as follows: "Whereas a variance theory explains a behavior or a characteristic of an object, a process theory explains the pairing or other rearrangement of mutually autonomous objects, such as the bets of the players and the number on the roulette wheel, whose individual courses are determined independently of one another by forces external to the core of the theory." A variance theory can explain, by identifying the necessary and sufficient conditions, why at a particular moment in time an innovation is adopted by a particular firm. A process theory can explain, given that there is an innovation and that there are potential users, the event that there are potential users that adopt the innovation.

Mohr holds that process theory is not a watered down version of variance theory, in which a probabilistic element has been slipped in to make up for the lack of identified sufficient conditions. Process theory explains specific types of events in a particular way. The fact that you did not win in roulette can be explained by analysing the speed of the ball and the wheel, the shape and the weight, and so on. The same fact is also explained by taking into account that there were necessary conditions (the bet, the ball and the wheel), and a chance of one in 37 of winning. For most purposes this is an entirely satisfactory explanation of the event of not winning.

What type of theory would constitute a satisfactory explanation of innovation diffusion, a variance theory or a process theory? Gerybatze describes the situation in diffusion research as follows: "Nicht überwunden wurde aber (...) die Vernachlässigung des dynamischen, prozessmässigen Charakters der Diffusion. Nahezu alle Diffusionstudien bauten auf einem Paradigma auf, das als Varianztheorie bezeichnet wird. Es wurden Korrellationen zwischen zahlreichen ökonomischen, psychologischen, organisationalen und sozialen Variabelen einerseits und Merkmalen der 'Innovativität' von Individuen und Organisationen andererseits ermittelt. (...) Dennoch weiss man aber nicht, wie diese vielen Faktoren sich im Zeitablauf gegenseitig beeinflussen und den Verlauf der Diffusion prägen. Es mangelt an einer Prozesstheorie der Diffusion (...). Dieser Mangel ist besonders gravierend deshalb, weil die Diffusion ein kausal- genetischer Prozess ist, bei dessen Analyse varianz-theoretische Konzepte versagen (...)." [Gerybatze (1982, p.231)]. The diffusion process, according to Gerybatze, propagates its own causation, has its internal coherence and logic. A process theoretical explanation of diffusion would illuminate this coherence of the sequence of adoptions, something which is difficult to accomplish by variance theory. This gain in understanding of the coherence of the process, however, goes at the expense of tractability down to the individual decision maker.

Intuitively, Gerybatze seems right in pointing out that there is a causal relationship between innovation adoptions that follow upon each other, either directly or indirectly. Diffusion seems to be a process with a certain internal logic and regularity. However, the extent to which adoptions can be explained by previous adoptions, relative to other causes, is an empirical matter. There is no a priori answer to the question whether it is more fruitful to explain the spread of innovations in an economic system as a coherent process, or rather as a sequence of not directly related adoptions. It is not clear a priori that process theories will succeed where variance theories seem to fail. In subsequent chapters, the introduction of innovations in firms will be analysed using both perspectives. We shall look at both adoption and diffusion models, employing variance and process theories respectively.

2.3.2 The introduction and spread of innovations

Diffusion processes have aroused the interest of social scientists from different backgrounds. There is a considerable tradition in diffusion research among sociologists [notably Rogers (1983)], anthropologists and geographers [e.g. Hägerstrand (1956)]. Attention of economists to this phenomenon dates back to seminal articles of Griliches (1957) and Mansfield (1961). There are some differences in approach between different disciplines. Whereas other social scientists have usually been concerned with the spatial patterns of diffusion, economists commonly deal mainly with the time dimension.[6] Also, the aspects of innovation diffusion that attracted the attention of sociologists are different from those studied by economists. Sociological research seems to concentrate on the mechanisms, whereas economic research primarily looks at outcomes of the diffusion process. The economic research in the field is predominantly occupied with the identification of factors, notably cost and benefit variables, that correlate with the diffusion speed. What is often considered less important, is an analysis of the events that actually take place in the course of the diffusion process, in terms of interacting, forming expectations, searching for information, decision making, adjusting and learning. This contrasts with the sociological approach of Rogers. He concentrates on social mechanisms, on communication networks, decision procedures, public opinion, change agents, authority, norms, etcetera [Rogers (1983)]. His analysis centres more on social processes and less on the inputs and the outcomes.

Rogers defines innovation diffusion as the spread in time of an innovation through certain communication channels among the members of a social system. There are three explanatory variables in the diffusion process that figure in this definition: the innovation, the communication channels and the social system. To put the issues involved in perspective, it is useful to consider

[6] An exception would be Karlsson (1988). His study combines time and spatial aspects of diffusion.

some aspects of each of these elements. We shall look at the appropriability of the innovation, the use of market channels as channels of communication, and the existence of market institutions and property rights as important ingredients of the social system.

An innovation is the first commercial use of an invention. Innovations can be categorised according to their character as: 1) product innovations; 2) process innovations; 3) market innovations; 4) organizational innovations; 5) legal and institutional innovations. In practice different types of innovations are often related. Process innovations frequently require new organizational arrangements or stimulate the development of new products. Product innovations stimulate the exploration of new markets and the development of new production processes.

This study deals predominantly with process innovations, but this is not to be interpreted too narrowly. Introducing a new process can involve adjustment of the organization of labour, an upgrading of the final product, an innovation in marketing, everything that improves efficiency in production of value. A process innovation is thus often accompanied by other changes and can stimulate further developments, e.g. new features and improved product design.

According to Rogers, innovations have the following five attributes:

1. Relative advantage: the degree to which an innovation is perceived as better than the idea it supersedes.
2. Compatibility: the degree to which an innovation is perceived as being consistent with the existing values, past experiences, and needs of potential adopters.
3. Complexity: the degree to which an innovation is perceived as difficult to understand and use.
4. Trialability: the degree to which an innovation may be experimented with on a limited basis.
5. Observability: the degree to which the results of an innovation are visible to others.

"In general, innovations that are perceived by receivers as having greater relative advantage, compatibility, trialability, observability, and less complexity will be adopted more rapidly than other innovations." [Rogers (1983, pp. 15-16)]. Although Rogers was not writing with process innovations in firms in mind, his attributes are applicable to some extent to these innovations too. Relative advantage is generally measured in terms of costs and benefits, in terms of efficiency. The most important attribute of a process innovation to an adopting firm is that it saves costs and increases productivity or capacity. But the other attributes are also forceful factors influencing adoption decisions. Compatibility and complexity determine the costs of adjustment when some innovation is to be introduced. The more compatible and the less complex an innovation, the less costs have to be incurred, first on research and assessment, and after the firm has decided to adopt, on organizational change and training. Trialability and observability determine the risk of adopting an innovation. The more triable an innovation, the less commitment is required to gain experience. Greater observability of an innovation reduces the costs of information gathering. Summarizing Rogers' partitioning, it can be said that there are two essential aspects to an innovation, the relative advantage in terms of costs and benefits that it can bring in the real sphere, and the difficulties that it involves to the control sphere: to assess the advantages and risks of adoption, to learn to work with it and to adjust working practices. Although relative advantage is a necessary condition and an important explanatory factor for a successful diffusion of an innovation, it is often not sufficient to explain adoption or non-adoption. Costs connected with compatibility, complexity, trialability and observability can speed up or retard adoption.

Innovations can take different physical forms. An innovation can be contained in knowledge or in a source of information.[7] One could think of something like the chemical formula for a drug or a blueprint of a machine. It can also be embodied in a consumption or investment good. Furthermore, an innovation can be tied to labour (as treated in the literature on learning by doing), to workers that have developed some skill or knowledge which is not costlessly transferable to others. An example could be a good software programmer, who has developed skills specific to some problem area. Finally, an innovation could be contained in an organization, in its structure, its routines or culture. It then has the character of an intangible asset. The physical form of an innovation determines its degree of appropriability. An innovation in the form of ideas, recipes and blueprints is less appropriable than an innovation hidden in a machine or tied up in an organization. The degree of appropriability of an innovation determines the channels through which it spreads, the costs of the transfer and the time it takes for an innovation to diffuse. Innovations in forms that permit property rights to be effectuated and protected can spread through market channels. The trade in a new machine embodying an innovation is governed by the same laws of demand and supply that govern any other market process. Innovations whose transfer involves high transaction costs, like complex knowledge, spread *through schools and courses. Innovations that have a more accessible form spread through other* channels: there is search for information, reverse engineering and imitation. Summarizing one can state:

1. Information and knowledge are usually cheap to reproduce. It is often hard to enforce exclusive ownership on disembodied knowledge or information. According to Baldwin and Scott, "Severe, even crippling, sources of market failure have been identified, when knowledge, in and of itself, is viewed as the commodity transferred or diffused in a process of voluntary exchange. If such abstract knowledge is envisaged as salable, its marginal cost of production approximates zero." [Baldwin and Scott (1987, p.114)]. However, information and knowledge are sometimes expensive to transfer, because it may take a lot of education and training to be able to make use of new knowledge.
2. New commodities, consumer and capital goods, are reproduced in a traditional production process and transferred by selling. Production capacity limits the availability of an innovation. Property rights can be enforced more effectively on commodities that embody an innovation than on knowledge *per se*. However, patent evasion and reverse engineering can constitute problems.
3. The reproduction and transfer of skills is a matter of teaching and learning. Learning capacity limits the adoption speed of an innovation. The diffusion of skills is partly regulated through the market for schooling. The relevant price is not only the sum transferred in the market, but also the opportunity costs of the time and effort spent by the trainee.
4. The reproduction and transfer of an organizational innovation also depends on learning and on factors like the flexibility of job contracts and organizational hierarchies.

[7] Metcalfe (1988, p.563) distinguishes between technology as knowledge and technology as artefact.

> Intangible assets are difficult to transfer, but organizational change is to some extent tradable, and its spread tend to be propelled by management consultants and business schools.

In connection to these issues, Baldwin and Scott distinguish between two different types of diffusion process. On the one hand, diffusion can be initiated or encouraged by the innovator. This type of diffusion is also called dissemination. One can differentiate between vertical and horizontal dissemination [Baldwin and Scott (1987, p.117)]. The mechanisms that lead to vertical dissemination are the same as those leading to purchases of raw materials or sales of final goods. Horizontal dissemination frequently takes the form of allowing another firm to produce a patented or otherwise proprietary product under licence. If unauthorised copying of licensed technology is effectively prevented, the innovator thus loses none of the monopoly control over its knowledge, but merely chooses whether to reap the monopoly gains through profits on its own sales or through the royalty fees charged its licensees.

On the other hand, diffusion can be without the permission or approval of the innovator, initiated by other agents. This type of diffusion is coined imitation. "Unauthorized imitation is a major diffusion mechanism when patents are easily circumvented, when high litigation costs and uncertainties make patents little more than a 'licence to sue', and when 'reverse engineering', or the analysis of how a competitor's product was made, is routinely pursued." [Baldwin and Scott (1987, p.120)]. The appropriability of the innovation determines the type of likely diffusion. If property rights can be enforced, then a producer of an innovation will probably market his product. If property rights are limited or cannot be secured, imitation is prone to occur.

The mainstream of diffusion literature focuses on vertical dissemination, looking at sellers and users of an innovation as clearly different groups and assuming them to deal through market channels. Diffusion literature has little to say on horizontal dissemination, which involves not only the sale of the innovation but also affects the licensor's competitive position on his own regular product market. A lot of research addresses licensing, but usually not to the extent that conclusions about the (optimal) time profile of the spread of an innovation can be drawn [Baldwin and Scott (1987, pp. 118-119)]. Imitation as a strategic alternative to innovation has been considered by several researchers, but not in connection to the diffusion rate [e.g. Martin (1984), Lieberman and Montgomery (1988)].

If the channel through which an innovation spreads over adopters is a market, then it is feasible to apply the usual market models to innovation diffusion. There can be an influence of market structure on diffusion rates, and there can be distortions of the market mechanism and market failure. The relationship between innovative activity and the structure of the product market has been the subject of extensive research. The usual Schumpeterian conclusion is that some degree of monopoly or monopolistic competition is a necessary condition for dynamic efficiency [Kamien and Schwartz (1982), Baldwin and Scott (1987)]. Empirical evidence, however, is mixed [Scherer (1980), Cohen and Levin (1989)]. The relationship between, on the one hand, investment in new technology and time of adoption and, on the other hand, firm size and market concentration has been extensively researched [e.g. Mansfield (1968), Davies (1979)]. Empirical evidence on the relationship between concentration and diffusion speed is mixed [Baldwin and Scott (1987, p.132)].

A market functions efficiently under a number of preconditions, one being full and free information about the commodity being traded to all parties involved. One market distortion is introduced by the problem of reduced appropriability. A related problem emanates from the fact that asymmetries of information are unavoidable in the trade of knowledge. A full disclosure of information on the innovation is not viable, if the innovation is contained in that very piece of information. Obviously, if this information is part of the traded commodity, secrecy must interfere with the market process. Information available to the traders in the market is incomplete, asymmetrically distributed over supply and demand, but probably expanding over time as diffusion advances. Information is incomplete, because commonly there is considerable uncertainty about both the technical and the economic characteristics of an innovation, about performance and profitability. This uncertainty hampers the estimation of the value of the innovation and thus the determination of the demand curve. In the words of Baldwin and Scott: "There is a paradox in the market for knowledge. Markets are presumed to work efficiently only if buyers and sellers have complete and accurate information on the commodities being traded. But by definition, this necessary condition for efficiency is violated, since buyers of knowledge do not know what they are acquiring, nor what its worth to them will be, until after they have bought it. For if they had sufficient knowledge, they would not have to acquire it."

Furthermore, the supplier of the innovation often has substantial market power vis-a-vis buyers, since he is much better informed than the potential buyer. Often the supplier is not willing to reveal all information about the innovation to the market, since this would eliminate the reason to pay for this information and thus destroy the market for the innovation. However, the opposite asymmetry is also likely to occur: "Alternatively, the discoverer of a new unit of knowledge (such as an independent scientist) may not be in a strong position to exploit it, and thus the preponderance of market power may lie with large enterprises which are the main potential customers of a new technology." [Baldwin and Scott (1987, p.115)]. Finally, experience with the innovation builds up in the course of time. This decreases the risk involved in determining the value of the innovation and in introducing it. The collection of experience depends on both trialability and on observability of the innovation.

2.3.3 Implications for model construction

These considerations about the process of innovation diffusion and the relationship between adoption and diffusion have some implications for model building:

1. If successive adoption decisions are largely unrelated, then microeconomic adoption models, together with a distribution of decisive characteristics over firms, fully explain diffusion of innovations. Regularity in diffusion patterns is then caused by some regularity in the distribution of some firm characteristic. If successive adoptions are strongly related to each other, then diffusion may be adequately explained by a model at the aggregate level. Then a detailed inquiry into the characteristics of individual firms might not add much to our understanding of the diffusion process. In this case microeconomic underpinning of diffusion models may be superfluous, since each adoption is mainly determined by previous adoptions by other firms, by a process at a higher level of aggregation than the particular firm. It may also be a fruitless thing to explore, because the relationship between micro characteristics and diffusion patterns is weak.
2. Adoption of innovations is partly a real sphere process and partly a control sphere process. The timing of adoption of an innovation is not only determined by relative advantage, but also by necessary investments in search, work adjustment and training,

which depend on observability, complexity, trialability and compatibility.

3. Diffusion of new technology is only partly a market process. There are problems of appropriability and, related to this, problems of disclosure of information. There can be considerable transaction costs. This limits the applicability of standard models of market behaviour.
4. To the extent that diffusion is accomplished outside the market, e.g. by free transfer of information or by costless imitation, models of innovation diffusion cannot rely on cost benefit arguments as explanation for observed patterns. If diffusion happens by imitation, then it is not so much constrained by profitability of the innovation, but by capacity to gather and sort out information.

2.4 Some distinctions in approaches to modelling technological change

The related processes of adoption and diffusion have inspired a large research effort over the last decade or so. Numerous studies have been conducted, both theoretical and empirical, after the mechanisms of the spread of new technology. There are several excellent surveys of the literature available [Stoneman (1983, 1986, 1987), Thirtle and Ruttan (1987), Baldwin and Scott (1987)] and a number of critical assessments of the state of the art [Gold (1981), Metcalfe (1988), Dosi (1991), Silverberg (1991)]. Rather than to repeat their work, I shall single out a couple of issues in the rest of this chapter which are of importance for the construction of models of technical change, and on which different researchers have made different choices. The purpose is to illuminate the choices that have been made in subsequent chapters, and to put them into perspective. Before tackling these issues, some pages will be used to review two models in more detail: a model of induced innovation and the so called epidemic diffusion model. The first describes the development or adoption of innovations at the level of the firm, and the second at the level of the industry. The first model can be classified as a variance theoretical model and the as second a process theoretical model. The first explains a change in technology as a logical consequence of optimizing behaviour and technical constraints. The second explains a change in technology as a chance process, where the changes in the circumstances determine the changes in the probability that a firm will adopt a new technology. These models are outlined here, because they can be used as illustrations later in this chapter, and more importantly, because they will serve as points of departure for model constructing in later chapters.

2.4.1 Induced innovation

The theory of induced innovation is an attempt to relate the rate and direction of technical change to the structure of factor demand and changes therein. The original formulation of the general idea is due to Hicks: "A change in the relative prices of the factors of production is in itself a spur to invention, and to invention of a particular kind - directed to economising the use of a factor which has become relatively expensive." [Hicks (1932)]. This assumption, that the direction of technical change is in part determined by factor cost ratios, or changes therein, has been analysed by a variety of authors during the 1960's [see e.g. Kennedy (1964), Dandrakis and Phelps (1965), Samuelson (1965)]. They dealt with the question whether a rise in the wage rate provokes labour saving innovations which could lead to unemployment. The theory has been contested by Salter, who claimed that technological change has no intrinsic factor saving bias, because firms attempt to reduce all their costs, no matter what they are spent on [Salter (1966)]. In his model, the marginal revenue of every factor is equalized in equilibrium, and the firm is indifferent between marginal reductions in factor demands of any kind.

Most authors have studied the conjecture of induced innovation in a macroeconomic context. Also they generally assumed a static framework: decisions on innovation are taken with only the present state of the system, characterized by factor prices and an innovation possibility frontier, in mind. Moreover, the focus in this work is usually on the determinants of the direction of technical change, assuming the rate of progress fixed exogenously. Binswanger and Ruttan have developed the theory further, giving it a content based in microeconomics [Binswanger and Ruttan (1978)]. Few authors have used the idea of induced technical change in a microeconomic model of the firm [see e.g. Kamien and Schwartz (1969), Magat (1979), Sato and Suzawa (1983), Sato and Ramachandran (1987)]. The latter models referred to are elaborations of the contribution of Kamien and Schwartz (1969). Kamien and Schwartz analyse a model which is microeconomic, dynamic and in which both the rate and the direction of technical change are decision variables. The model describes the intertemporal decision problem of a firm maximizing discounted profit. By investing in factor augmentation, the firm can determine both direction and rate of technical progress. Their analysis confirms Hicks neutrality in equilibrium as the optimum for the firm.

The firm models in subsequent chapters draw heavily on the approach of Kamien and Schwartz. In the next chapter, a similar type of model will be considered, introducing however a number of alternative assumptions concerning the firm's production possibilities and its decision rules. Some new elements will be introduced in line with the conclusions drawn in the last two sections above. The technical features of the model will be retained, however, which allows for analysis of the model by standard methods and for the derivation of some elegant results.

In order to focus on the effects of technical change and the optimal rate of factor augmentation for a profit maximizing firm, Kamien and Schwartz (1969) posit very simple assumptions about the environment. The firm operates in a static and certain world. It faces a stationary demand curve and constant factor prices which are unaffected by the firm's course of action. The production function is of the neoclassical type:

$$\phi(F(BK,AL)) = y \tag{1}$$

Here K and L are capital and labour respectively. B and A are 'augmentation' parameters, such that BK and AL are amounts of capital and labour in efficiency units. F is a continuously differentiable linear homogeneous function, determining the curvature of the isoquants. The function $\phi(F)$ is monotone increasing in F and $\phi(0) = 0$. The transformation ϕ affects only the spacing of the isoquants, not the curvature. Capital and labour are assumed freely and continuously variable in unlimited quantities at constant costs of r and w respectively. The firm can substitute between labour and capital along the isoquant without costs. It can also adjust the scale of production to the optimum without adjustment costs.

When B and A grow at different rates, this affects the curvature of the isoquant. This will change the optimal ratio of capital and labour. It can be shown that the relative change in the optimal capital labour ratio over time depends on the relative changes in B and A, together with the elasticity of substitution. Thus by constructing an optimal time path for B and A, an optimal time path for the capital labour ratio follows. The properties of ϕ assure that there is an optimal scale of production. Thus the net revenue function depends exclusively on choices concerning factor augmentation. Choices for the growth rate of B and A determine time paths for K and L, and for production y.

Factor augmentation is assumed to proceed at a certain cost and there is a trade off between capital and labour augmentation. For a fixed budget, it is assumed that there exists an innovation possibility frontier, as in Kennedy (1964). Let a dot over a variable indicate a time derivative. For one ***particular*** rate of spending on technical advance $\bar{M}$, the maximal proportional rates of factor augmentation which may be achieved are related through $\dot{A}/A = g(\dot{B}/B) = g(\beta)$, where β is defined by the last expression and where:

$$\beta \geq 0; \qquad g(\beta) \geq 0; \qquad g'(\beta) < 0; \qquad g''(\beta) < 0 \tag{2}$$

Feasible combinations of proportional rates of factor augmentation are thus related a by downward sloping concave function $g(\beta)$, lying entirely in the first quadrant. Selection of a value for β determines $g(\beta)$ and thereby the direction of progress $g(\beta)/\beta$. In addition, it is supposed that more spending on technical progress can shift the innovation possibility frontier outward. The relationship between the expenditure of an amount M and the position of the frontier of feasible rates of factor augmentation is expressed by the function $h(M)$. The function $h(M)$ is a nonnegative, monotone increasing, concave function of M:

$$M \geq 0; \qquad h(M) \geq 0; \qquad h'(M) > 0; \qquad h''(M) < 0 \tag{3}$$

Putting the above elements together (scaling such that $h(\bar{M}) = 1$) the expressions determining factor augmentation are now as follows:

$$\dot{A}/A = g(\beta)h(M) \qquad \dot{B}/B = \beta h(M) \tag{4}$$

The firm maximizes the sum of discounted net revenues, subject to the constraints on the improvement of technology. Given a demand curve for its products which is unaffected by the behaviour of the firm, constant factor prices and a production function, net revenue N is a function of B and A only. Let ρ be the constant discount rate, then the firm's object function is written as:

$$\text{Max}\, Z = \int_0^\infty e^{-\rho t}(N(A,B) - M)dt \tag{5}$$

subject to:

$$\dot{A} = Ag(\beta)h(M) \qquad \dot{B} = B\beta h(M) \tag{6}$$

The model can be analysed using the Maximum Principle of Pontryagin. In order that the values β and M, for all time t, be optimal, there should exist continuous functions λ_1 and λ_2, such that β and M provide a maximum at every point in time t for the Hamiltonian:

$$H = e^{-\rho t}(N(A,B) - M) + \lambda_1 A g(\beta)h(M) + \lambda_2 B\beta h(M) \tag{7}$$

The transversality conditions must hold, implying that value of the costate variables λ_i should approach zero as t goes to infinity. It is assumed that the discount factor ρ is large enough to ensure convergence of (5) and the existence of an interior solution.

To determine β and *M*, the Hamiltonian is differentiated with respect to these two variables and the derivatives are set to zero. This yields:

$$g'(\beta) = -\frac{B\lambda_2}{A\lambda_1} \tag{8}$$

$$h'(M) = \frac{e^{-\rho t}}{A\lambda_1 g(\beta) + B\lambda_2\beta} \tag{9}$$

The time derivatives of the costate variables are obtained by differentiating the Hamiltonian with respect to the state variables *A* and *B*:

$$\frac{\delta H}{\delta A} = e^{-\rho t}\frac{\delta N}{\delta A} + \lambda_1 g(\beta)h(M) = -\dot{\lambda}_1 \tag{10}$$

$$\frac{\delta H}{\delta B} = e^{-\rho t}\frac{\delta N}{\delta B} + \lambda_2\beta h(M) = -\dot{\lambda}_2 \tag{11}$$

Using (4) and rearranging terms, this can be written as:

$$A\dot{\lambda}_1 + \dot{A}\lambda_1 = -e^{-\rho t}A\,\frac{\delta N}{\delta A} = -e^{-\rho t}wL \tag{12}$$

$$B\dot{\lambda}_2 + \dot{B}\lambda_2 = -e^{-\rho t}B\,\frac{\delta N}{\delta B} = -e^{-\rho t}rK \tag{13}$$

Kamien and Schwartz prove the second equalities above to hold (Lemma 5, op. cit., p.676). Equations (12) and (13) say that in equilibrium, marginal benefits from factor augmentation equal marginal costs. Thus, given that shadow prices of technical change λ_i go to zero as t goes to infinity, these equations yield on integrating:

$$A\lambda_1 = \int_t^\infty e^{-\rho s}wL\,ds \tag{14}$$

$$B\lambda_2 = \int_t^\infty e^{-\rho s}rK\,ds \tag{15}$$

These expressions are convenient, because the costate variables are expressed in terms of quantities and prices only. We can substitute equations (14) and (15) into (8) and (9) respectively. Equilibrium values of β and *M* must satisfy at all times *t*:

$$g'(\beta) = -\frac{\int_t^\infty e^{-\rho s} rK ds}{\int_t^\infty e^{-\rho s} wL ds} \tag{16}$$

$$h'(M) = \frac{e^{-\rho t}}{g(\beta)\int_t^\infty e^{-\rho s} wL ds + \beta \int_t^\infty e^{-\rho s} rK ds} \tag{17}$$

Thus we have equilibrium expressions for the slope of the functions *g* and *h*. The changes in these slopes over time can be traced, and since the general shape of the curves *g* and *h* is known from conditions (2) and (3), the changes in slope indicate the direction of the changes in the optimal values of β and *M* themselves. Therefore the time derivatives of (16) and (17) are considered:

$$\frac{dg'(\beta)}{dt} = \frac{e^{-\rho t}(rK + wLg'(\beta))}{\int_t^\infty e^{-\rho s} wL ds} \tag{18}$$

$$\frac{dh'(M)}{dt} = h'(M)\{h'(M)(rK\beta + wLg(\beta)) - \rho\} \tag{19}$$

From (18) it follows that β will reach an equilibrium value when:

$$g'(\beta) = -\frac{rK}{wL} \tag{20}$$

The stationarity of β requires $g'(\beta)$ to be stationary. According to (20), $g'(\beta)$ is stationary when rK/wL is constant. This is the case when *K* and *L* grow at the same rate. Kamien and Schwartz prove (Lemma 4, op. cit., p.671) that in equilibrium the growth differential of *K* and *L* is proportional to the growth differential of *A* and *B*. Thus *A* and *B* are required to grow at the same rate. This implies that in long run equilibrium β equals $g(\beta)$. Thus the equilibrium direction of technical change is Hicks neutral and is independent of the long term factor price ratio [cf. Kennedy (1964) and Samuelson (1965), which generate a similar outcome in a macroeconomic framework].

The long run development of the optimal *M* can grow or to decline monotonously, depending on the development of total costs. Kamien and Schwartz then continue to consider the stability properties for the optima for β and *M*. These turn out to be dependent on the elasticity of substitution and on the cost function.

The result of Hicks neutral technical change in equilibrium is not very surprising. It is assumed in this model that the two factors of production, capital and labour, have exactly the same properties, and only differ in their prices. Capital and labour appear symmetrically in the production function, they can be substituted for each other without cost, and they can be purchased and scrapped without limit against a fixed price at any moment in time. To this model a mechanism of induced technical change is added that takes care of compensating the factor price difference by changing the productivity of the factors. The relatively more expensive factor is made relatively more productive. If the elasticity of substitution is smaller than one (Kamien and Schwartz (1969), p. 680), the relatively expensive factor will then be used relatively less. This decreases the revenues from further investments in augmentation of that factor. The process continues up to the moment that further augmentation of both factors yields an equal increase in revenue. Then the symmetry between capital and labour is complete: factor augmentation of both factors proceeds at the same rate.

2.4.2 The epidemic diffusion model

The second model to be outlined here is the most prominent and widely used model in the field of diffusion studies: the logistic or epidemic model. This model has been introduced in economic analysis over three decades ago [Griliches (1957)], and has in several modified forms been employed extensively since then [notably by Mansfield (1961, 1968), Nabseth and Ray (1974)]. The mathematical representation of the basic model is very simple:

$$\dot{n}_t = \beta n_t \left(1 - \frac{n_t}{N}\right) \qquad (21)$$

Here n_t is the number of potential adopters of the innovation, who have already adopted at time t, a dot indicates a time derivative, β is a parameter measuring the steepness of the diffusion curve, the speed of the process, and N is the total number of potential adopters of the innovation which is assumed to be constant in time. The model has been used in various fields. One application stems from epidemiology where it is used as a description of the spread of a contagious disease over a homogeneous population. Suppose that there is a population that has N individuals, of which n_t are ill at time t. The chance that one individual meets any other is β. The chance on contagion of the disease equals the number of individuals that are ill, times the chance that they meet another individual, times the probability that this other individual is healthy. Together this gives equation (21). This differential equation can be solved:

$$n_t = \frac{N}{1 + \exp(-\alpha - \beta t)} \qquad (22)$$

Equation (22) is a sigmoid curve which has three parameters: β determines the slope, α the point at which the curve begins, and N the ceiling that will be reached. A lot of empirical work has sought to explain variation in these parameters over different diffusion processes, by relating them to all kinds of economic variables, like profits, costs, cash flow, liquidity, etcetera. The justification of the use of the logistic curve in economics has ranged from stressing some analogy between the spread of epidemics and innovations to labelling the curve as a mere "summary of the data" [Griliches (1957)]. A sophisticated use of the logistic model in a Schumpeterian framework is presented by Iwai (1984). Iwai considers the appearance of a series of innovations in the course of time and a chance mechanism to describe imitation, assuming that "the probability that a firm is able to copy a particular production method is proportional to the frequency of firms which employ that method in the period in question." [Iwai (1984), p.165].

The limitations of the model are generally recognized, however, and have been most thoroughly analysed by Gold (1981), but also by others [Stoneman (1982), Freeman (1988), Coombs, Saviotti and Walsh (1987), Rosegger (1986), Thirtle and Ruttan (1987)]. The main criticisms are of two kinds. First of all, the model lacks a decision theoretical basis: it "pushes aside many of the more interesting theoretical questions - such as the nature of the adoption decision by the individual enterprise - and substitutes a rather mechanistic hypothesis of behaviour." [Karlsson (1988)]. Secondly, the model is rather rigid: post innovation improvements are left out of the picture; the number of potential adopters is fixed; the group of potential adopters is homogeneous.

2.5 Issues in modelling the introduction of new technology

The model of Kamien and Schwartz and the logistic diffusion model originate from two different research traditions. The literature on models of induced innovation is firmly rooted in the neoclassical tradition, with an emphasis on theoretical analysis of profit maximizing behaviour under constraints. The Kamien and Schwartz model describes decision making behaviour of a single firm, trying to optimize its future income stream by investing in improvement of its technology. Technology is thereby identified with factor productivity. The model does not specify explicitly whether this investment is in R&D, in ready to use technologies developed outside the firm, in adapting technologies supplied by capital producers or in imitating competitors. It only assumes that investment raises productivity and that there is a choice in both extent of productivity improvement and mix over factors.

The literature on epidemic diffusion models is more empirically oriented, with less emphasis on the analysis of firm behaviour and more eye for friction in the spread of information and for adjustment mechanisms that operate in disequilibrium. The focal point of the analysis is a qualitatively defined technology. The logistic diffusion model describes the changes in the level of use of a production technique, an aggregate variable, leaving the decision behaviour of the individual firm implicit. In this respect, it is similar to most models that describe a phenomenon on the aggregate level that results from a variety of activities on the micro level.

The two models describe different aspects of the same phenomenon: the introduction of new technology. The logistic model summarizes the outcome of this process; the induced innovation model describes in general terms the optimal investment behaviour that could be underneath. These two models appear next to each other here, not because one is to be presented explicitly as the decision theoretic basis of the other, but because they constitute two ways of looking at technological change in an industry. The spread of new technology in an industry involves two dimensions: over *time* a variety of *firms* in an industry go through a range of *technological levels*. Thus there is variation over time in the technology dimension and in the firm dimension, of which either model consider only one. The epidemic model picks out a technological level and considers the number of firms that reach or pass this level over time. The model of induced innovation considers a single firm and describes how it passes through different technological levels over time. Both approaches to the description of the introduction of new technology have some illuminating features, and both have their blind spots. The rest of this chapter will be used to shed some more light on some of these aspects, also in reference to other models.

2.5.1 Equilibrium or disequilibrium

Traditional explanations of economic development often assume that an economy tends to progress along some type of long term equilibrium growth path, occasionally deviating temporarily from this track in response to frictional imperfections. There are factors like time lags in the adaptation of the capital stock, interest inelasticity of investment demand, limited factor substitution and wage and price inflexibilities, which might lead to departures from equilibrium growth. These factors are all of a technical nature, pointing at physical and organizational limitations in the economic system, which hamper equilibrium growth, despite the tendencies of rational agents to move towards an equilibrium.

Evolutionary economists, however, stress that the equilibrium concept by itself fails to capture the essential characteristics of economic development. Economic dynamics are not to be described as an equilibrium process, not so much because of the above frictional imperfections, but, first of all, because of the existence of genuine uncertainty, and secondly, because of features of economic behaviour itself: entrepreneurial behaviour by nature is an attempt to disrupt equilibrium in search for new profits opportunities. Technological change is a central ingredient to this view: it is both a source of this fundamental uncertainty and it is an important element in competition, as already pointed out by Schumpeter. "In Schumpeter's framework it is disequilibrium, dynamic competition (in the sense of 'imperfect' competition) among entrepreneurs, primarily in terms of industrial innovation, which forms the basis of economic development. Thus, the emphasis is on the supply side, that is, autonomous investments rather than on demand induced accelerator investments or multiplier processes (demand push) as driving forces in economic development." [Freeman, Clark, Soete (1982), p.31].

This issue turns up, not only in studies on innovation, but also in research into technology diffusion. In the context of the diffusion process, equilibrium and disequilibrium are terms applying to dynamics at the aggregate level. Disequilibrium can refer to different things. First of all, it may mean a deviation from equilibrium, some sort of stationary situation, due to adjustment lags or delays in the spread of information. Metcalfe relates disequilibrium to imperfect information and to adjustment time. In his words, the issue is "... whether diffusion is to be viewed in terms of an equilibrium or a disequilibrium process (Griliches, 1957), whether diffusion patterns reflect a sequence of shifting equilibria in which agents are fully adjusted and informed, or whether, by contrast, they reflect a sequence of imperfectly perceived disequilibria lagging behind the development of a 'final' equilibrium position." [Metcalfe (1988), p.561]. Disequilibrium development results when agents are not fully informed about and adjusted to the changing circumstances. The use of the disequilibrium concept in this sense applied to diffusion is to a certain extent arbitrary. The spread of information and the process of adjustment of agents may be called an equilibrium process itself. As Metcalfe notes in this respect: "... one can turn any disequilibrium model into an equilibrium equivalent and vice versa by a suitable definition of the information sets and perceptions of adopting agents." [Metcalfe (1988), p.561].

Secondly, disequilibrium can mean a deviation from equilibrium, in the sense that there is an inconsistency between planned behaviour and the diffusion process as it develops. There may be two types of inconsistency. On the one hand, planned behaviour of a firm at some point in time may be inconsistent with optimal behaviour at a later moment, thus leading to revisions of plans over the diffusion process. This may occur in the diffusion process if adoptions of a new technology generates new information that induces revisions of plans. On the other hand, firms' planned courses of behaviour may be mutually inconsistent. Mutual inconsistency of plans may occur in the diffusion process if information about competitors is incomplete. Note, however, that disequilibrium at the aggregate level does not have to be at odds with rational or optimizing behaviour at the level of the individual.

Dosi hints at the latter type of disquilibrium as he stresses the importance of the flow of information which is produced and released during the diffusion process. In his view, feedback loops from adoptions of innovations to further diffusion and further innovative activity are central to the understanding of the process of technical change. He refers to diffusion dynamics as an equilibrium processes, "whenever micro decisions are postulated to be *reciprocally consistent* and 'rational' microbehaviors all turn out to be fulfilled in their objectives". Conversely, disequilibrium diffusion processes are "all those dynamics wherein (a) the 'attractors' of the process change themselves as a result of the very actions of the agents - such as when there are system-level increasing returns to technology adoption and/or (b) the diffusion process is explicitly represented in terms of the trial-and-error efforts of the agents, which exhibit 'disequilibrium behaviors' and deliver 'disequilibrium signals' to other agents." (Dosi, 1991, p.191).

The disequilibrium between individual plans and the resulting process of diffusion that Dosi refers to is a consequence of particularly strong feedbacks, like the unforeseen emergence of increasing returns, which account for a certain type of unpredictability of the future course of the process: "Of course, if *positive feedbacks* between adoption of innovations, their further improvements, and the cost of acquiring them are important enough, this implies, in technical terms, a source of non-convexity in the technological opportunities at the level of the firm, the industry or the whole economy. We leave the world of *convergence* to macro-level *solutions* and of equilibrium paths which can be defined independently of the *actual* technological and economic history of particular clusters of innovations. On the contrary, we are in the path dependent world of Arthur (1983 and 1988) and David (1985), wherein the long term positions of the system may well depend on even minor initial fluctuations, individual choices, institutions, and policy measures." (Dosi, 1991, pp.198-199). Disequilibrium in diffusion, in Dosi's view, is thus rooted in unpredictable externalities of innovation adoptions with such a pervasive effect that they lead to radically different conditions in the further course of the process.

Finally, disequilibrium may imply a by-passing or a rejection of the notion of an equilibrium path of diffusion as a point of reference altogether. This assertion may be related to a rejection of the assumption of optimizing behaviour. Dosi seems to hint at such a view when referring to 'trial-and-error efforts of agents' and 'disequilibrium behaviors'. Nevertheless, disequilibrium at the aggregate level in this sense, based on microeconomic 'non-optimizing' behaviour, does not preclude a certain regularity in the pattern of diffusion, e.g. because the market works as a selection mechanism.

For the purpose of modelling, the first description of disequilibrium does not clarify a lot, as already indicated. The second description, inconsistency of individual plans and aggregate outcomes, seems an intuitively relevant characterization of disequilibrium in diffusion. It hinges on the assumption of incomplete, heterogeneous and dynamically changing sets of information. The problem is how to capture those incomplete, heterogeneous and dynamically changing information sets in a model. One may distinguish between real unpredictability and limited use of information. Numerous models are based on the assumption that agents do not make full use of all available information represented in the model (e.g. money illusion and adaptive instead of rational expectations), which may lead to inconsistency of agents' plans and periods of adjustment. This is not fully satisfactory in modelling diffusion, since it leaves out the important but unpredictable feedback loops from adoptions to innovations. Yet, the possibilities to deal with these unpredictable factors are much smaller than those to model limited use of information. In the next chapter, a firm model of induced innovation, based on the Kamien and Schwartz model above, will be used to explore (disequilibrium) aggregate behaviour, where firms employ heterogeneous information sets and make individually optimal but mutually inconsistent plans. Lastly, the third description of disequilibrium applies in a certain sense to the epidemic diffusion model, in the sense that no explicit reference is made to equilibrium. The course of diffusion is expected to follow a regular pattern, which is asserted without being based on optimizing behaviour or any other behavioural rule.

2.5.2 Exogenously or endogenously propelled diffusion

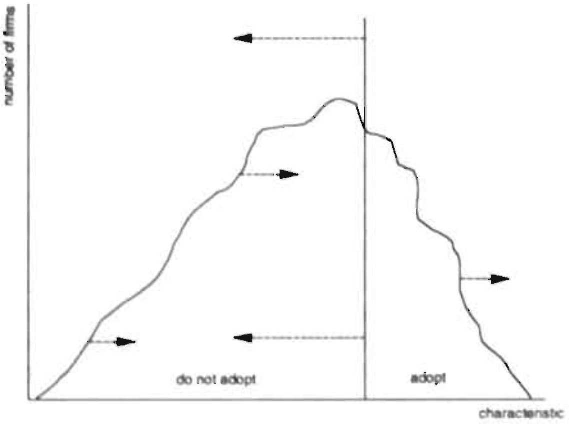
Diffusion of innovations and technical progress in firms are processes in time. The sequence of events that makes up such a process can be comprehended in two ways. One can assume that the process is driven by external factors, or that it is propelled endogenously. If the process is driven by external causes, no event in the sequence is causally related to other events in the sequence, the order of events is exogenous, time is exogenous, and so are the speed of the process and the serial correlation of events. If a process in time is understood as a sequence of events without internal cohesion, then its state at any moment in time can be described by a static model. If the process is driven endogenously then there are feedback loops between the process as it progresses and its own further development. In such a dynamic process, speed is endogenous and the events have an endogenously determined sequence. To describe such a process one needs to model the way every next event depends upon its predecessors; a description of the state at any one moment has to take the history of the process into account.

A type of model where the diffusion process is described as a sequence of events, driven by causes which may be either exogenous or endogenous to the diffusion process itself, is the probit diffusion model [see e.g. David (1969, 1975), Davies (1979)]. The main assumptions of this model are, 1) that there is a heterogeneous population of firms; 2) that a firm adopts the innovation as soon as benefits exceed costs; 3) that there is full information on benefits and costs; and 4) that gains change over time and/or costs of adoption change over time. These changes in gains from and costs of adoption are both exogenous to the diffusion process in the probit model. David describes the general features of the model as follows [(1969), quoted by Reinganum (1989), p. 894]:

> "Whenever or wherever some stimulus variate takes on a value exceeding a critical level, the subject of stimulation responds by instantly determining to adopt the innovation in question. The reasons such decisions are not arrived at simultaneously by the entire population of potential adopters lies in the fact that at any given point of time either the "stimulus variate" or the "critical level" required to elicit an adoption is described by a distribution of values, and not a unique value appropriate to all members of the population. Hence, at any point in time following the advent of an innovation, the critical response level has been surpassed only in the cases of some among the whole population of potential adopters. Through some exogenous or endogenous process, however, the relative position of the stimulus variate and the critical response level are altered as time passes, bringing a growing proportion of the population across the "threshold" into the group of actual users of the innovation."

Potential users of the innovation are assumed to differ in some crucial aspect, say size. If there were positive returns to scale in using the innovation, then adoption at high prices would be profitable for the larger firms only. Costs of innovation adoption, however, could change e.g. because of increased competition or technical progress in the innovation supplying industry, because of changes in wages or prices of inputs, but also as a consequence if the diffusion process itself. As the price of the innovation would drop, adoption would gradually become attractive for ever smaller firms. Suppose that a density function *f*(*size*) describes the distribution of the population of firms over different sizes. As adoption costs fall over time, or as firms grow, a diffusion curve emerges (see Figure 2). Off course, the model may be refined by assuming the adoption to influence profitability, and thereby growth. Thus the size distribution may change endogenously. Beside size, the crucial characteristic determining the profitability of adoption can be many things, but in the literature a couple of other variables have been singled out for analysis: search costs, age structure of the capital stock [Karlsson (1988), p.23], prior beliefs on uncertain benefits, transport costs and factor costs [David and Olsen (1986)], and risk-aversion.

Figure 2: The probit model



The arrows indicate the movement of the curves over time.

Generally two different elements can be encountered in the literature which are assumed to explain the adoption of an innovation: changes in available information about the innovation and changes in profitability of the innovation. In the probit models mentioned above, changes in profitability moves diffusion forward. Exogenous changes in information drive diffusion in a model by Jensen [Jensen (1982)]. He assumes new information about the innovation to be released every period. The information has a certain probability of being positive, if the innovation is truly profitable, and a lower probability of being positive, if the innovation is in fact unprofitable. The firm starts out with a prior belief about the qualities of the innovation and adjusts its beliefs as new information accumulates. On the basis of its new expectations, the firm can either decide to adopt the innovation or to wait for the next piece of information. It is shown that it is rational for the firm to adopt the innovation, as the expected probability of profitability exceeds a threshold level. The diffusion pattern then depends on the initial distribution of prior beliefs about the innovation.

The models reviewed here contain a number of elements that are appealing in the explanation of diffusion. The decision to adopt depends on perceived costs and benefits, and variation in adoption times is caused by the fact that there is a variety in firm characteristics. However, there seems to be a limitation to this explanation of diffusion, in so far as the 'driver' of the process remains outside the model. As Metcalfe states, stressing the spread of information: "The crucial point here is that the information sets of the population of potential adopters are independent of the number of actual adopters of the innovation. They may well change for other reasons but these are unrelated to the adoption process *per se*." [Metcalfe (1988), pp.564-565]. The process of diffusion itself does not cause the generation of more or better information on the innovation. Agents are fully informed about the innovation the moment it comes on the market. Thus fashion and bandwagon effects, learning effects and reduction of uncertainty are not included as an explanatory factor for diffusion. Moreover, not only a feedback loop from previous adoptions to the current availability of information on an innovation may be important, but also feedbacks from adoptions to profitability, possibly through other variables, like market shares and growth, prices of inputs and outputs, supply of the innovation or incremental innovations. Therefore, to model diffusion as driven by external causes might be to miss a substantial part of the process.

2.5.3 More on endogenously driven diffusion

Treatments that are explicitly concerned with feedback loops of the diffusion process to further adoptions have emanated from game theory and industrial economics. This work has concentrated predominantly on identifying characteristics of firms, innovations and markets which are likely to affect incentives to adopt, but has not had a significant impact on applied work. Two types of feedbacks from adoptions can be distinguished: effects of adoptions on information about the innovation, and effects on profitability of the innovation. In models dealing with the first type, a mechanism of expectation formation or learning is presupposed. Expectations may be defined as a set of possible outcomes, together with a probability distribution connected to these outcomes. Learning is in this context the modification of expectations on the basis of experience. In the present context, expectations are notions about the properties of a new technique that can be formed before or after adoption. There will be expectations, on the one hand, on the operating price, adjustment costs, performance and profitability of the innovation and, on the other hand, on the further development of the new technology, both the possibility of improvements and the chance of replacement by a more efficient successor [see e.g. Rosenberg (1976)]. Firms can base their expectations on their own adoption of an innovation or on the adoptions of others. In the latter case there are technology spill-overs.

Endogenous expectation formation on the technical quality of an innovation has been explored in Stoneman (1981). There a learning mechanism is specified in a model describing the behaviour of a single firm which every period adjusts its expectations concerning the performance of the new production technology and accordingly chooses a 'portfolio' of production techniques [Stoneman (1981)]. The firm has an idea of profitability of the technique and an estimation of the risk involved. Every period the firm uses the technology, experience build up, and the firm adjusts its estimate of expected profit and variance accordingly, following the rules of Bayesian statistics. According to Stoneman, "This is the closest we get to an 'economic' theory of learning." [Stoneman (1981)]. The main elements of the model are a utility function that depends positively on expected profit and negatively on the anticipated variance of the returns (as agents are assumed to be risk averse), a Bayesian learning mechanism and quadratic adjustment costs. Together, a choice of technique procedure, learning and adjustment costs produce an intra-firm diffusion path.

In models dealing with the second case of feedback loop, profitability of adoption is assumed to vary with the number of earlier adopters. Profitability can vary in two directions: it may decrease or increase as firms are later in the row of adopters. The first, falling profits, is likely if early adoption leads to above normal profits, but further diffusion of the innovation puts a downward pressure on product prices. The second, rising profits, may occur if there are network externalities connected to adoption. An example of the first is the work of Reinganum [Reinganum (1981)]. She models the user side of the market for an innovation and abstracts from the producer side. The innovation is a cost-reducing, capital-embodied process innovation, ex ante firms are homogeneous, and firms have perfect information. Like in the probit model, firms adopt as soon as gains from adoption exceed costs, and costs of adoption are assumed to decrease exogenously. The gain from adopting decreases as the firm is further down the adoption queue. It can be shown that under specific conditions on the costs of adjustment, strategic behaviour can lead *ex ante* identical firms to adopt a new technology sequentially. In effect, in every period a number of firms adopts, until the gains in profit equal the costs of adoption. The remaining firms wait until the next period, when costs have decreased. This model assumes an exogenous propagator of the diffusion process on the supply side of the market, but explicitly models the influence of adoption on the profits of all firms in the industry. Thus, the mechanism that stops adoptions, before full diffusion has been reached, is endogenous. Firms take each others decisions into account when deciding to adopt, thereby capturing an important part of the dynamics of diffusion. Stoneman considers this work particularly Schumpeterian in spirit, a feature which he describes as follows: "The conception is that an entrepreneur innovates and the attractiveness of attaining a similarly increased profit and the pressures on the costs of the old technologies in a new regime encourage others to imitate, this imitation representing a diffusion process." [Stoneman (1986, section 2)].

Another way of endogenizing the fall of profit rates leading to further adoptions of innovations is to explicitly represent the role of the supply side in the diffusion process. Changes in the interaction of demand and supply lead to changes in prices and profitability, which again induce changes in quantities demanded and supplied. The papers of Metcalfe (1981), Bass (1978) and Glaister (1974), Stoneman and Ireland (1983), David and Olsen (1986) and Ireland and Stoneman (1986) explicitly deal with the supply side in diffusion. The supply curve of the innovation may shift over the course of the diffusion process, on the one hand, because it is likely that there are decreases in costs of producing the innovation, as the producing firm goes down the learning curve, and on the other hand, because supply may well be concentrated at the start of the diffusion process but the market structure is likely to change from monopoly or oligopoly towards increased competition as the innovation spreads and the threat of entry increases. A description of the supply of innovations has been combined with some form of the epidemic model for the demand side by Glaister, by Bass and by Metcalfe. Glaister makes the diffusion speed parameter in the logistic model dependent on the price of the innovation and Metcalfe makes the size of the total population of potential adopters price dependent. Glaister shows that, when monopoly supply is associated with an epidemic demand model, the predictions of the model change and the logistic curve no longer appears. Metcalfe shows that in his model the logistic curve can be retained, however, only at the expense of assuming rising costs of producing the innovation. Metcalfe's model is rather mechanistic in the sense that capacity expansion is a fixed percentage of returns and is not explained as the result of rational firm behaviour. David and Olsen (1986) considers a probit model, assuming competitive supply of the innovation and learning on the part of the innovation suppliers. This latter feature causes an endogenous fall in the supply price of the innovation, inducing diffusion to proceed. Stoneman and Ireland (1983) also adds a supply side to probit models. They model both monopoly and oligopoly supply of the innovation. This allows them to endogenize the sales price of the new product, using the assumptions of profit maximising suppliers and cost reduction as experience accumulates.

Both exogenous and endogenous forces drive the model of Ireland and Stoneman (1986), in which they explore different types of expectations formation within the context of the probit model [Ireland and Stoneman (1986), see also Stoneman (1987)]. On top of the probit structure, where firms adopt a new technology once some characteristic surpasses a critical threshold, different expectations formation regimes are imposed. Following Rosenberg, they distinguish between expectations on price and on technological performance. The latter concerns the exogenously given chance of obsolescence of the innovation. The former type is modelled as myopic and adaptive expectations, consistent over or under estimation and perfect foresight. Thus the price of the innovation changes endogenously over the diffusion process. The expectations on the price movements are predetermined in the case of perfect foresight and develop endogenously in the other cases. Two types of supply side are added to the model, monopolistic and oligopolistic supply. Differences in diffusion patterns and welfare implications are derived.

In the models above, every next adoption of the innovation exerted a negative externality upon the other firms in the industry, by putting profitability under pressure. In specific cases, however, innovation adoptions confer positive externalities upon other adopters. In such cases, a slow diffusion is explained by the relative unattractiveness to adopt early. The profitability of an innovation increases with its use, if there are network externalities. Katz and Shapiro (1985) distinguish between three possible sources of network externalities. Firstly, externalities may be generated through a direct physical effect of the number of purchasers on the quality of the product (e.g. telephone or communication systems). Secondly, externalities may be generated through an indirect effect of the number of purchasers on the quality of the product. This phenomenon may be present when a good (hardware) needs compatible complementary inputs (software) to put it to use (e.g. the number of computer or cd-player users has an influence on the availability and quality of software resp. compact discs). Thirdly, externalities may arise when the quality and availability of post-purchase service for the good depend on the experience and size of the service network, which may in turn vary with the number of units of the goods that have been sold (e.g. a particular brand of car).

An important factor influencing the size of a network, and thus the scope of the positive externalities, is the extent of standardization. Illustrative cases of efficiency losses, due to lack of standardization, are to be found in computer technology, video equipment and car telephones. A number of authors have dealt with aspects of this problem. Arthur (1989) considers competing technologies and the possibility of lock-in; David (1985) describes the history of the rather early standardization of the typewriter keyboard, leading to lock-in; Katz and Shapiro (1985) consider the effect of consumption externalities on competition and the form of the market equilibrium and deal with the compatibility decision of the firm; Katz and Shapiro (1986) consider the technology adoption decision of firms when there are two competing technologies.

Returning to the models of section 4, it may be noted that both the epidemic diffusion model and the induced innovation model of Kamien and Schwartz are dynamic models, in the sense that in both models the time path, the order of events and the speed of the process which is described, is endogenously determined. In the epidemic model, there is a clear feedback from adoptions to further adoptions. As more agents adopt the innovation, experience grows, information spreads and uncertainty diminishes. The number of adoptions in any time period, or equivalently, the chance that any one firm adopts in any period of time, is a function of earlier adoptions. The reason for diffusion is usually assumed to be the information feedback from further adoptions. In the Kamien and Schwartz model, there is a feedback loop of another type. Every period there is investment in technical progress M in a direction characterized by $g(\beta)/\beta$, and the effect of this investment on capital and labour productivity is related to all progress in the past on which it improves, and to all progress in the future for which it will be the basis. Note that in the epidemic model, the feedback is an externality, whereas it is not in the Kamien and Schwartz model. In this last model, the environment is stationary and the firm takes account of all effects from future adoptions of innovations in its current planning.

2.5.4 Optimizing or following routines

Whether one regards diffusion as a disequilibrium process, in one of the senses distinguished above, or not, may be related to how the underlying individual agent's economic behaviour is understood. The decision about adoption or non-adoption of new technologies is a complicated matter, in which many cognitive processes and procedures play a role (as has been elaborated in section 2 above). What is needed for the construction of a model with a microeconomic foundation is a representation of the basic elements of this choice process. Two stylized representations figure in the economic literature: behaviour understood as optimizing and as routinized. Optimizing behaviour is assumed in the game theoretic and probit models described above; institutionalized behaviour, though not necessarily non-optimizing behaviour, drives the traditional logistic models of Griliches (1957) and Mansfield (1968).

The distinction between optimizing and routinized behaviour is stressed by evolutionary theorists like Nelson and Winter (1982). The traditional assumption is that agents optimize under constraints. The constraints may be many and restrictive, but even if we abstract from part of them, it is assumed that we still end up with a good approximation of economic behaviour. Evolutionary theory questions the appropriateness of optimizing behaviour as a stylized representation of decision making, especially when modelling processes of technological change. It claims that the constraints on decision makers are so binding and restrictive, due to the complexity of technological progress, the fundamental uncertainty involved in this type of decision making, and the limited capacities of agents to oversee the consequences of their choices, that a description of behaviour as being guided by routines and rules, or by trial-and-error, comes closer to reality. From this perspective, the observed order at the aggregate level, the regularity of the diffusion path, is not produced by order at the micro level, but is the result of diversity at the micro level, of learning by individual agents, in combination with environmental selection. An environmental selection mechanism takes care of eliminating inefficient choices through competition. Models in which these routines and decision rules figure at the micro level can easily produce disequilibrium features, but under some conditions they yield robust patterns of development at the aggregate level.

The choice to describe decision behaviour as routinized rather than optimizing can be thought of as a matter of practical concern, and does not necessarily have to be a consequence of a fundamental difference in view on how agents decide. At issue is how much one should be concerned with modelling rational choice, if the subject is so restricted in his capacities to act optimally that it is likely that decisions will be guided by rules of thumb and guesses on the basis of experience. It might be more practical for purposes of description and prediction, not to bother too much about the motivation of the agent and his marginal profit calculations, but to model common strategies, vested routines and rules of thumb.

A diffusion model that models firm's behaviour as institutionalized along these lines is the simulation model of Silverberg, Dosi and Orsenigo: "Decision-making is incorporated on the one hand in certain robust rules of thumb (for the most part feedback rules dealing with oligopolistic pricing and production policies) and 'animal spirits' in the form of decision rules governing replacement policy (the payback period) and expansion of capacity ('estimates' or 'guesses' of future demand growth corrected by experience)." [Silverberg, Dosi and Orsenigo (1988), p.1037]. In their model a new technological trajectory is introduced. Firms differ in their evaluation of the prospects of this trajectory, notably the profitability of the new techniques that will be developed along this trajectory. To take advantage of the new technology in the future, it is important to build up the required skill level now. Thus, given expectations, it can be rational to invest in new technology before it satisfies the usual efficiency criteria. By means of computer simulations it is shown that, given divergent expectations, a rather chaotic pattern of adoption decisions and market share developments evolves, in which some firms win the competitive struggle, despite the fact that they were by no means first movers, nor last movers, in adopting the innovation. A remarkable feature of the model is that, despite the considerable range of microeconomic diversity and disequilibrium, at the aggregate level the S-shaped form of the diffusion curve stands out.

2.5.5 Discrete technologies or continuous technical change

Technological progress appears in a large variety of models, of which diffusion models and adoption models are but particular classes. Generally, diffusion models describe the diffusion of a specific innovation over a range of potential adopters. This innovation can be a new capital good, some new type of machinery which is acquired by purchase from some manufacturer, or it can be a new method of production that is developed in house. In both cases, the technology which appears in the model is discrete. It is an innovation, qualitatively different from earlier methods of production, pieces of machinery, or other. The innovation emerges on the market and in the industry as an alternative to the traditional technique. A recurrent theme in diffusion literature is the observation that in reality this sharp distinction between the new and the old does not exist [see e.g. Gold (1981), David and Olsen (1986), Dosi (1991), Silverberg (1991)]. Mostly a new technology is being developed further after its first launch on the market. The adopter who adopts first acquires something different than adopter who adopts last. Moreover, it is important to recognize that this process of incremental development of the innovation is not independent of the diffusion of the innovation. Innovation and diffusion are mutually dependent processes. Diffusion is a precondition for learning by doing and learning by using [Rosenberg (1982)], and thus contributes to the accumulation of experience necessary for further incremental innovations. This common occurrence of post innovation incremental improvements, the gradual development of a dominant design [Abernathy (1974)], has some implications for modelling diffusion. Firstly, there is no clear dichotomy between innovation and diffusion (as argued also in section 3). Diffusion entails further innovation on the part of both developers and users. Secondly, post innovation incremental improvements make that the ultimate scope of diffusion changes over the course of the process. The range of applications of the novelty grows over time, and therefore the number of potential adopters increases.

Whereas most diffusion models, amongst which the logistic model, assume a sharp qualitative distinction between the old and the new technology, and therefore confront us with the difficulties indicated, models which identify technology with factor productivity, like the models of induced innovation, are unable to distinguish between qualitatively different techniques. Thus, on the one hand, many diffusion models are restricted because they assume a simple dichotomy between the old and the new technique, and do not account for the interaction of diffusion and incremental innovation. On the other hand, production function models generally cannot illuminate adoption of innovations, because there is no clear distinction between qualitatively different techniques, no representation of qualitative barriers which have to be overcome when firms switch to a new type of technology. There are drawbacks to both a discrete and a continuous representation of technological advance. Attempts to find a solution to these can be found in Iwai (1984), Metcalfe (1988), Soete and Turner (1984) and in Diederen *et. al.* (1990). In these papers, a distribution of firms over a range of techniques is assumed, which is transformed during the diffusion process.

2.6 Conclusions

The purpose of this chapter has been to set the stage for the modelling exercises in the chapters to come. In the first part, a general representation of an economic system has been sketched, stressing the distinction between the real sphere of activity and the control sphere. It was argued that technological change is a product of cognitive activity, and to model technological change it is important to elaborate a representation of the possibilities of and constraints to activities in the control sphere. Limitations to cognitive capacity introduce bounded rationality and are a source of uncertainty. These are important determinants for the direction and speed of technological progress.

Furthermore, some aspects of the phenomena of innovation adoption and diffusion were highlighted. It was pointed out that innovations can have different characteristics, that they may be either disembodied or embodied in different forms, which leads to a variety in transfer mechanisms. Transfer may be through market channels or through other communication channels and may or may not be actively promoted.

In this chapter, two models which deal with the introduction of new technology, each from a different perspective, have been outlined: a model of induced innovation and the epidemic diffusion model. These models have been presented, because they will be referred to in later chapters. To put them in perspective, we reviewed a number of approaches towards the explanation and modelling of innovation dissemination, centred around four themes which have commanded an interest in recent literature. Firstly, diffusion can be qualified as an equilibrium or as a disequilibrium process; secondly, it can be seen as propelled by exogenous or by endogenous forces; thirdly, individual behaviour can be thought of as optimizing or as routinized; fourthly, innovations commonly undergo further development during their diffusion, and thanks to their diffusion. It was argued that disequilibrium in diffusion, understood as mutual inconsistency of firms' plans and deviations of realized diffusion from earlier plans, may be an important phenomenon, because of the interdependence of innovation adoptions and information generation. Possibilities to catch the feature of unpredictability of this information generation process in a model are limited, however. Furthermore, it was asserted that there are good arguments for the case that diffusion is driven at least partly by endogenous causes. Finally, it was argued that the choice to model economic behaviour as either optimizing or as routinized behaviour is a matter which should be decided in connection to the purpose of the modelling exercise. In matters of technology adoption, agents may be severely constrained in their attempts to optimize, e.g. by lack of information and restrictions in the capacity to evaluate information, such that their optimal decision rules work out to be rules of thumb and guesses on the basis of experience. Model complexity may then be reduced without much loss of predictive capacity if routines are assumed instead of optimization under constraints.

A recurrent theme has been the relationship between successive innovation adoptions: how and to what extent does one adoption influence the next. In particular, the question is whether there is something inherent to the process of diffusion that accounts for the empirical regularity of a (more or less) sigmoid diffusion curve. Some diffusion models are built on the assumption that the regularity of the diffusion curve is caused by factors exogenous to the diffusion process itself, e.g. by a regular (bell shaped) distribution pattern of decisive firm characteristics. Most models, however, allow for an element of feedback from early adoptions to later adoptions. Adoptions at present can influence later adoptions in different ways. A current adoption may help to release information, because its introduction starts a learning process and experience with the new technology is being built up. Moreover, a current adoption may affect the profitability of the innovation. Mostly the expected profitability of adopting is reduced as diffusion proceeds, but if there are positive network externalities, profitability may well go up with diffusion.

If the regularity of the diffusion curve is inherent to the diffusion process, because these feedback loops have a dominant influence on the course of the process, then one may wonder how important it is for understanding the aggregate pattern to model adoption at the firm level. Some evolutionary economists hold that regularity at the aggregate level is likely to result in spite of irregular and complex behaviour at the micro level, because competition functions as a selection mechanism *ex post*, rewarding some behaviours and putting a penalty on others. If diffusion along a specific pattern of curve is a robust phenomenon, compatible with different kinds of micro behaviour, then this might call for a 'process theoretical' explanation of diffusion, rather than a 'variance theoretical' explanation. An understanding of the probability mechanisms at the aggregate level might be more illuminating for understanding diffusion than a description of firm behaviour.

3. A model of firm behaviour

3.1 Introduction

It has been argued in the preceding chapters, that models of technological change developed along traditional lines are often based on restrictive premises regarding the availability and the costs of dealing with information. In particular, many models abstract from both genuine uncertainty and costs of information gathering and processing. Costliness of information gathering and processing implies that one cannot assume that all agents have the same information at their disposal upon which to base their decisions, and that there are costs to the decision making process itself.

It is likely that in the case of adoption of a new technology by a firm, abstracting from uncertainty and the costs of dealing with information leads to distortions in the representation of economic behaviour. The process of deciding on technological change is often lengthy and costly, and there is a certain degree of uncertainty involved. The decision to implement a new technology is typically taken stepwise, and involves costs along the way: costs of studies and pilot projects, learning and training costs, costs of capital investments and adjustment costs. As stressed in the vintage literature, firms do not change their production capital instantly, the moment a (qualitatively) new type of capital good becomes available. They gradually switch to the use of new capital goods, by investing every year in a new vintage of the new type of capital, and by scrapping some machinery of the old type. This means that the firm invests over many years in the change from one technology to the next, and may even start switching to its successor, before finishing the switch to a specific technology completely.

The models in this chapter describe the investment planning of a firm at a certain moment in time. A firm is assumed to maximize discounted future income, by using two instrument variables, size of investment and 'direction' of investment. The direction of investment is a measure for the ratio of capital deepening and capital widening, the ratio of investment in productivity gain and investment in capacity expansion. The firm is faced with constraints, which express its production possibilities and opportunities for technological change, and its restrictions in buying inputs and selling output. The solution to this maximization problem is a time path for the two instrument variables. The values for period one of this planned time path determine the current actions of the firm. The next period, the firm is assumed to repeat the same procedure, and to adjust its plans if circumstances have changed.

The firm cannot take events which are fundamentally uncertain into account in its planning, so the models cannot account for this aspect of the information problem (one could assume, however, a relationship between perceived degree of uncertainty and the discount rate). The firm assumes in its planning a certain regularity in future developments. The aspect of costliness of information, on the other hand, can be represented in models. It is supposed that the firm operates at a certain moment in time in a specific way, and that without a decision to change the present course, firm operations continue as before: the same production volumes will be produced with the same techniques. A change in operations only occurs, if the firm takes a decision to change. To do this, it has to invest in changing current routines. Thus there is no change without costs, and the larger the change, relative to the current practise, the higher the costs.

Information which is used in the process of planning is represented in the models as price information. The firm forms certain expectations about prices of in- and outputs, and draws up its investment plans accordingly. Technical information is treated as a commodity, as something that can be generated or bought and on which investment funds are spent. Thus, whereas current price information is freely available, technical information is considered proprietary and has a price. This is expressed by a sort of production function for technology. New technology is expressed as a gain in productivity of the firm, and this productivity gain in itself has a cost which is independent of the technology of competitors in the industry. This means that we abstract from technology spill-overs, bandwagon effects, costless imitation and learning from competitors. This is a limitation of the model, and implies a partial disregard of the interdependence of the technology in use in different firms over time, of the endogenous character of technical change in an industry. A representation of this aspect is deferred to chapter 6, where the perspective is shifted and technology itself will be the focal point of analysis.

The plan of the chapter is as follows. In the next section some additional assumptions on firm behaviour will be considered. The model of Kamien and Schwartz (1969), outlined in section 4.1 of chapter 2, will serve as point of reference. The assumptions upon which the Kamien and Schwartz model is constructed are amended, to fit the general framework presented so far. In the third section, the basic model of firm behaviour is introduced. This model differs from the Kamien and Schwartz model in specification and interpretation, but less in form. In contrast to Kamien and Schwartz, we shall assume that the firm is tied to its own history in determining its production volume and its production technique, and continuously invests in improvements, given its current starting point. The methods and instruments to examine the characteristics and implications of the model, used by Kamien and Schwartz, can still be applied. In the rest of the chapter, this altered model then serves as a tool, that will be gradually extended and used to examine the implications of our main assumptions on firm investment behaviour, on technical progress and on competition. First, the nature and effects of technical progress will be highlighted by contrasting the basic model with a model in which technical progress is absent (section 4). This contrast can be used to examine the effect of expectations concerning opportunities for technical progress for investment planning. Secondly, the concept of a technological trajectory, on which technical opportunities are gradually depleted as time progresses, will be incorporated in the model (section 5). Thirdly, the assumption of perfect competition is dropped and the possibility of rivalry on input and output markets, through a price mechanism, will be considered (section 6). Both the demand for the final product and the supply of the variable inputs will be assumed to have finite price elasticities. Finally the process of competition will be illustrated with the help of a simulation model (section 7).

3.2 Firm behaviour

We shall assume that the firm maximizes future discounted profits under constraints. This assumption is both very common and very general, and it seems warranted to spend some effort on finding a sensible interpretation. Various answers are possible in this context on questions like: what is a firm, whose are the future profits, what kind of constraints are relevant to the firm and what can a firm do to attain its objective.

3.2.1 The description of the firm

Consider the following characterization of a firm. A firm is a coalition of agents, related to each other on a contractual basis, with the purpose of producing for sale, and thereby to gain an income. It is a type of organization which exists because of the benefits to be reaped from division of labour in production. There are different types of agents involved in such an organization, of which share holders, employees and managers are the most important groups. There is a specific distribution of decision making power among the agents, which is largely regulated by contracts and by law. Generally the firm owns a number of assets, like capital goods, company buildings, brand names, patents, licences and so on. In any period of time, the firm buys additional inputs, like raw materials, energy and intermediaries, and uses those in combination with the services of its assets for production. Beside that, the firm can invest in its assets.

The inputs into production differ in several respects, but for company planning the degree of 'fixedness' or 'variableness' is an important dimension. For our purposes, an input is defined as variable, if it can be bought and sold on a current basis, if it can be substituted for without costs of disposal.[1] An input is defined as fixed, if it cannot be disposed of. Thus, in the present context, a factor of production is regarded as fixed, if its costs are sunk costs to the firm.[2] For our purposes we assume a strict separability of variable inputs and fixed factors of production.

In reality, factor inputs can be variable or fixed to different degrees. To what extent an input is variable, depends on the time period under consideration. What is (quasi-) fixed in the short run, can be variable in the long run. The degree of 'variableness' of a factor of production depends first of all on the costs that the firm would have to cover if it would change or dispose of them. The costs of alteration or disposal of an input of production depend on a number of factors. On the one hand, there are the adjustment costs in the usual sense, which depend on technical determinants and the existence of markets: there are costs of dismantling and selling; for dedicated equipment or production facilities there may be no market. On the other hand, there are costs which are determined by contracts and legal rules: labour is often a fixed factor, because of rights to tenure. A second, probably more important, determinant of the 'variableness' of a production input is the system of management and control within the firm, and the system of provision of information. A firm can be thought of as an organization in which the management is paid to take decisions on inputs and outputs, on the basis of all relevant information, in the interests of the shareholders. The interest of the shareholders would be the maximization of discounted future profits. However, from theories of managerial behaviour, from principal agent theory, it is known that managers may pursue a variety of goals beside profit maximization, and

[1] Consider electricity as an example: what is bought is also used.

[2] Since we abstract from the option to the firm to cease production activities altogether and exit, there is no need here to distinguish sunk costs from fixed costs along the lines of Baumol *et al.* (1982).

that information available to decision makers may be biased by those who provide it [see Mueller (1992) for an attack on the profit maximization hypothesis]. Also, as pointed out above, in matters of strategy and technology choice, there is likely to be such an abundance of relevant information, that not all information can be considered. Thus, whether an input is fixed or variable is not a technical datum, a given fact. Rather, managers *decide* which strategic options to consider, which technological choices to evaluate in depth, in short, which inputs to consider variable and which fixed.[3] They do this on the basis of imperfect and partial information, and maybe with other objectives in mind beside profit maximization.

Thus, for the sake of analysing the process of planning of activities in a specific time period by a firm, we shall describe a firm as an organization, endowed with a set of fixed factors of production (of which the costs are sunk), that maximizes future discounted profits by renting or buying variable factors for production: the firm *is* a set of fixed factors and *uses* variable factors. Which inputs are considered fixed for a certain period depends not only on adjustment costs, but also on a choice which is based on necessarily partial and imperfect information. The idea, that the first step in the firm's decision making process is to determine what is considered variable and what fixed, reflects the notion of bounded rationality. Bounded rationality can be understood as the a priori restriction of the dimensions of the search for an optimal choice (cf. section 2.1, chapter 2). Information gathering and evaluating has a cost, and the costs are likely to rise as more variables enter the decision making problem. Therefore, decision making processes may be described as proceeding in two stages. First, variables that will be taken into consideration are determined, and then, within the confines set by this a priori choice, an objective function is maximized.

Notice that, by considering the firm as made up out of fixed factors, it is put in historical time. The decision problem of the firm is largely determined by the fixed factors that are handed over from the past. The management does not so much decide on the amount of capital and labour they are going to employ, but on the way they are going to manage the firm as it already exists and use the capital and labour that they have got under their command. The capital outlays and labour force handed over from the past embody the firm's earning capacity, its value. The costs of these outlays are, under normal circumstances, sunk and do not enter the decision problem of management.

3.2.2 Some remarks on other descriptions of the firm

The implications for model construction of identifying the difference between fixed and variable inputs as the main distinction in inputs relevant for investment planning can be illuminated by recalling e.g. the model of Kamien and Schwartz (1969). There are two aspects of interest. First of all, some remarks can be made on the relevance of the capital labour distinction. In production a variety of inputs is used: equipment, buildings, raw materials, intermediaries, labour, energy, etc. Within this diverse group of inputs, a distinction can be made between 'primary factors' and 'other inputs'.[4] Productive factors, like labour and equipment, contribute to the production process and earn a wage or rent for rendering productive services. Other inputs, like raw

[3] Machines, buildings, whole production plants including workforce, may be considered either fixed or variable, depending on the circumstances. Under the threat of bankruptcy variables otherwise fixed may be considered variable.

[4] Other inputs are flow variables; productive services of primary factors are also flows; primary factors are stocks: the capital stock, the labour stock and other asset stocks.

materials, intermediaries and energy, are included physically in the final product or used up in the production process. They do not render services and do not earn an income, but are purchased and disappear as such in the process of production.

Production in a firm can be represented either as production of output, using various inputs, or as production of value added, using productive factors. In macroeconomic models of a closed economy, even when there are intermediary deliveries, the two representations are equivalent: if every input is produced somewhere in the economy, total production of final goods is equal to total value added. Total value added is divided among factors that deliver productive services. This is known as the circular flow: the total revenue of production is distributed among the productive factors in the economy, all payments are income to somebody. If one models the planning options of a single firm, however, the two representations are different. To the firm, the category of inputs that are used up in the production process do matter. There can be substitution or technical progress in the use of these inputs, and therefore these inputs should appear in the decision problem of the firm. If the inputs like raw materials, intermediaries and energy, are treated as given, the production function is restricted to be a function of productive factors only, and firm production is just modelled as value added, then one omits part of the (opportunities for) technical change.[5]

Within the category of productive factors, it is common to aggregate every factor into either the labour or the capital category, like in the Kamien and Schwartz model. This is a fruitful categorization, if one analyses matters of income distribution. A dichotomy between capital or labour is useful, if income from capital lands up with other agents, and is spent in another way, than income from labour. From considering demand for capital relative to that for labour, one can make inferences about the determination and the stability of the income distribution, the division of revenue between workers and capital owners, and its effects on economic development. If one models the investment planning of the firm, however, the distinction between capital and labour may not be appropriate to explain the decisions of the firm. To the firm it is total costs of inputs that matter for profits, not whether expenses are wages, interest payments or costs of raw materials. By investing in changes of production, the firm attempts to economize on every type of cost, irrespective of its source [compare Salter (1966), pp. 43-44].[6] In summary, in a model of a firm's investment planning, it is inappropriate to restrict production to mean value added, such that only productive factors are considered as inputs, and it not obvious why the firm would let its decisions be guided by aggregation of inputs into the two categories capital and labour.

A second point to be made here concerns the consequences of distinguishing fixed and variable inputs for the flexibility of the firm. In the model of Kamien and Schwartz, the firm is pictured as an organization with a large flexibility. Its only constraint is a production function, which describes the minimum required combinations of the inputs capital and labour, to produce a specific amount of output. The firm is in no way constrained to a certain amount of output or input of some kind. Every moment in time, the firm rents the optimal volume of capital and labour inputs, to produce the optimal amount of output. As prices change, the firm instantly and

[5] To assume implicitly, that these inputs can be included into either the labour or the capital category, seems also unsatisfactory. These other inputs can be a major part of the costs of the firm and there is no reason to assume that they have prices that move in line with wages or with the rate of interest.

[6] There is a distinction between an *investment*, which is a payment or a commitment now, in exchange for productive services over a prolonged period to come, and a regular *cost*, which is a current payment in exchange for current productive services or inputs. An investment can lead to fixed costs in the future, and wages can often be seen as regular costs, but the two are by no means the same.

costlessly alters its operations. The firm is flexible in two senses: there is costless and immediate *adjustment of scale* of operations and there is costless and immediate *substitution* between the two inputs capital and labour along isoquants. Using this assumption of flexibility of the firm, the unconstrained substitution and changing of scale, it is generally difficult to explain inter firm differences in size, a competitive market with increasing returns, and slow diffusion of technology, without *ad hoc* auxiliary assumptions. Identifying a considerable part of the factors of a firm as fixed assets implies that the flexibility of the firm is limited by its own past. Adjustments in scale become costly, because changes in the capacity of the fixed factors are required. Substitution between variable and fixed factors is only possible in one direction, fixed for variable, and requires investment. Thus, in a model of a firm's investment planning, it may be more appropriate to assume that the firm has limited room for manoeuvre with respect to output growth, because it is tied to its past, and limited possibilities for substituting inputs, because part of the inputs are fixed. Moreover, changes in scale of output or factor ratios require investments, and therefore time.

Summarizing the argument, one can state that the profitable employment of the fixed factors is the goal of the enterprise and the use of variable factors is instrumental to the attainment of this goal. The distinction between fixed and variable factors of production seems to be a most relevant categorisation of production inputs, because the firm treats these categories differently in working out an investment plan. Fixed factors are given, as they result from past operations. Variable factors are employed to help the (owners of the) fixed factors to an income. The categorisation fixed versus variable is not parallel to the dichotomy capital versus labour, since: 1) the larger part of labour and capital usually fall in the same category: they are fixed; 2) the larger part of the variable factors, raw materials, energy, intermediate goods, are not included in either capital or labour; 3) consequently, costs of variable factors are not the same as the wage sum and costs of fixed factors are not the same as interest payments; 4) fixed factors are considered fixed, in the sense that their costs are sunk. Therefore the firm cannot substitute variable factors for fixed factors, only fixed for variable factors.

3.2.3 The instruments of the firm and the technical constraints

Given the above characterization of the objective of the firm, now consider the instruments at its disposal. For simplicity, assume that there is a stock of fixed assets, which require a specific amount of variable inputs to be fully employed in production. In other words, there are fixed technical coefficients and there is a fixed production capacity. Assume that there is no depreciation nor technical obsolescence, that the firm sticks to its single homogeneous product and that there is no under-utilization of capacity.[7] Under these circumstances, the management of the firm can only do two things to pursue its objective. It can augment the production capacity of the firm or it can raise the productivity of the variable factors, thereby decreasing the variable costs per unit of product. The firm can expand or rationalize, invest in capital widening or in

[7] There are no quantity restrictions on product and factor markets and the marginal costs of producing at full capacity do not exceed the output price.

capital deepening. The application of both these instruments has its price in the form of investment. Given an amount of investment, there is a trade-off between using it for expansion and for rationalization. Any investment adds to the stock of fixed factors.[8]

Technical constraints determine the effects of investment expenditures. These constraints can allow for technical progress. Technical change in production function models is usually represented by a move *of* the isoquants. Substitution is a move *along* an isoquant and scale adjustment is a move *from* one isoquant *to* another. Kamien and Schwartz assume in their model that the first move is costly and the last two are for free and occur instantaneously upon price changes. If one assumes, that the firm's technical constraint is not a neoclassical production function, and that substitution and expansion cannot be free of costs, then the distinction between technical change and other activities that result in a larger production capacity or in changes in technical coefficients needs to be defined otherwise. A definition of technological change is required, other than 'a shift of the isoquants'.

Both in a situation with and without technical change, the firm can expand production capacity and raise variable factor productivity.[9] Without technical change, capacity can be expanded by enlarging the capital stock, adding another production line of the same machinery that is already there. Expansion means doing more of the same. By contrast, in a situation with technical change, expansion of capacity tomorrow does not occur in the same way as it happens today. With technical progress, expanding gets cheaper over time, e.g. because equipment decreases in price relative to its performance. Technical change implies that there is learning, that there is a cumulative effect. The starting point for the technology of next period is the technology, the knowledge and the skills of this period.[10] For the change in variable factor productivity, an analogous story applies. Without technical change, variable factor productivity can be raised by substituting fixed factors for variable factors, e.g. installing some device in a machine to make it more fuel efficient. This investment might be repeated in the same way on all machines the firm uses. With technical change, a cumulative effect comes in again: e.g. an energy saving method of production is developed and used with all equipment in the firm. Thus investment necessary to decrease variable factor requirements by the same volume would decrease over time. The future replacement of variable factors is cheaper than the present.

[8] This description of the firm is related to Porter's ideas on competitive behaviour: "In a static view of competition, a nation's factors of production are fixed. Firms deploy them in the industries where they will produce the greatest return. In actual competition, the essential character is innovation and change. Instead of being limited to passively shifting resources to where the returns are greatest, the real issue is how firms increase the returns available through new products and processes. *Instead of simply maximizing within fixed constraints, the question is how firms can gain competitive advantage from changing the constraints.* Instead of only deploying a fixed pool of factors of production, a more important issue is how firms and nations improve the quality of factors, raise the productivity with which they are utilized, and create new ones." [Porter (1990a), p. 21, italics added].

[9] Consider the following as an image. Given technological opportunities, investment in technology generates (disembodied) '*blueprints*' for production improvements. As current production capacity is larger, opportunities to exploit these blueprints is larger. Thus there are economies of scale and a 'cumulative effect'. Without technological opportunities, investment is in '*machinery*'. The extent to which the result of such investment can be used is independent from already installed capacity. Therefore there is no cumulative effect; there are no scale economies to this type of investment.

[10] The point could also be made as follows, using common symbols: $Y = \alpha K$, where Y is production, α is capital productivity and K is capital. Let I be investment and a hat over a variable indicate a relative change. Expansion without technical progress is an *absolute* increase in the capital stock, $Y = \alpha(K + I)$, and (capital augmenting) technical change is a *relative* increase in capital productivity, $Y = \alpha(1 + \hat{\alpha})K$. Given a price of capital, expansion through investment I in extra machines has a fixed price. If a relative change in capital productivity $\hat{\alpha}$ also has a fixed price, then expansion in absolute volumes through technical change gets cheaper over time. In the Kamien and Schwartz model, labour augmenting and capital augmenting technical change both have a fixed price.

In summary, both in an economy without and in one with technical change, investment is required to arrive at a larger output capacity or at lower requirements of variable factors. The difference between the two regimes is the occurrence of a cumulative effect, of dynamic economies of scale. If there is *no* technical change, the technical constraints to the decision problem in the next period are the same as in this period (abstracting from the possibility that opportunities for substitution can run out). If there *is* technological change, then capacity expansion and raising of productivity get cheaper over time. This happens, because the current expansion of capacity takes advantage of learning from previous expansions. Investment goes into a productivity rise of the capital stock, putting the firm in a better position to realize future productivity rises. Technological change is thus a learning process in time, that can proceed at different speeds, but that takes a course in which no steps can be skipped. This characterization of technical change implies that technical change is only a meaningful concept in a dynamic framework. In a comparative static framework expansion and substitution in absence of technical change cannot be distinguished from expansion and substitution involving technical change.

So altogether two different regimes can be distinguished in this context, stable and progressing technology, and under each regime two different choice directions are open to the firm, economizing on variable input requirements per unit of output and expanding output capacity. Given progressing technology, one can distinguish between fixed factor (or stock) augmenting technical change and variable factor (or flow) saving technical change. Given stable technology, one can distinguish between plain expansion and plain substitution.

3.2.4 Conclusion

The preceding discussion can be summarized in a number of statements:

1. For the analysis of the investment planning problem of the firm, the most relevant distinction between groups of inputs is between fixed and variable inputs. No method has been proposed above about the way total inputs are divided into these two categories, only some factors which might be of influence have been indicated. Given this division, the firm is assumed to optimize an objective function.
2. Abstracting from obsolescence and assuming positive profit opportunities at current output levels, there are two courses of action open to the firm: expansion and rationalization.
3. For any change in present operations, be it the expansion of production or the decrease of variable input requirements, investments are required.
4. If there is technical change, investments have a cumulative effect.

With these things in mind, we now turn to the Kamien and Schwartz model, presented in section 4.1 of chapter 2. Their model is specified in accordance with the fourth point: factor productivities rise as a consequence of investment and this results in higher factor productivities to start with in the next period. However, the model does not accord with the first three points. Firstly, the difference between fixed and variable factors of production is not accounted for: in their model all factors are variable. Secondly, the firm has also two basic instruments, but different ones than those proposed here: the firm is principally occupied with dividing its resources between producing a higher capital productivity and producing a higher labour productivity. Thirdly, the firm expands, contracts and substitutes between capital and labour automatically, freely and at once.

3.3 A basic model of firm behaviour

Assume that there is a firm in existence at time $t = 0$. At this moment, the firm is characterized by a certain output capacity Y_0 and variable factor requirement V_0, determined by its present stock of fixed factors and variable factor productivity. Assume that product prices and factor prices are constant and given (this assumption will be relaxed in section 6 below). Given opportunities for technical change, the management draws up a plan at time zero for future investments in enlargement of capacity and in raising variable factor productivity.

A firm's plan can be biased in either direction. On the one hand, a plan can be geared predominantly toward capacity expansion through investments in new production equipment, which embodies new technologies. These capital investments, given a certain variable factor productivity, will in general require additional variable inputs to be purchased or hired. This can require investments in training and reorganization, which are then complementary to the instalment of new capital goods. On the other hand, a plan can be aimed mostly at saving variable factors, by putting the accent on training of employees, streamlining of the organization, raising efficiency in the use of resources. Investments in new tools and equipment are then instrumental in attaining a higher variable factor productivity. Given real world technological opportunities, expansion and rationalization usually go hand in hand. In general, any expenditure plan affects both plant capacity and variable factor productivity, but all feasible plans will usually be biased in either one or the other direction.

Let the size of total investment in value terms at time t be M_t (or equivalently, let the volume be M and the price be unity and constant over time) and let the function $h(M_t)$ determine the effect of investment on expansion and variable factor productivity change. More investment will generate larger effects, but the marginal effect of investment on expansion and efficiency improvement is assumed to diminish. Thus we assume that $h(M)$ is upward sloping and concave [compare the Kamien and Schwartz model in section 4.1 of chapter 2]:

$$M \geq 0; \qquad h(M) \geq 0; \qquad h(0) = 0; \qquad h'(M) > 0; \qquad h''(M) < 0 \tag{1}$$

For convenience, the time index t of these variables has been suppressed.

For every amount of investment, the firm can invest in both expansion of capacity and in improvement of efficiency, such that a choice for more of one implies less of the other: there is a trade-off between expansion and rationalization. We assume that this trade-off can be represented by a Kennedy-type technological progress frontier, a function $g(\beta)$, such that for the value of M for which $h(M) = 1$, if the rate of expansion at time t would be chosen to be β_t, the maximum possible rate of productivity improvement would be $g(\beta_t)$. The ratio of $g(\beta)$ and β, of productivity improvement and expansion, can thus be called the direction of progress of the firm. The restrictions on $g(\beta)$ are the same as in the model of Kamien and Schwartz, i.e. $g(\beta)$ is downward sloping and concave:

$$\beta \geq 0; \qquad g(\beta) \geq 0; \qquad g'(\beta) < 0; \qquad g''(\beta) < 0 \tag{2}$$

Let production at time t be expressed by Y_t and variable factor demand at time t by V_t. The price of output is called P, the price of variable inputs w, and the constant discount or interest rate is ρ. Assume that capital markets are perfect and that a constant interest rate is expected. Let the present period be indicated by 0 and let the planning horizon of the firm be period T. The problem of the firm is to choose an optimal path for the amount and direction of investment, such that the present value of future income is maximized. Future income equals future cash flows, minus the yields that can be invested in the firm at a higher rate of return than the interest rate:

$$\max_{\beta,M} Z = \int_0^T e^{-\rho t}(PY_t - wV_t - M)dt \tag{3}$$

subject to:

$$\dot{Y}_t = Y_t \beta h(M) \tag{4}$$

$$\frac{\dot{\gamma}_t}{\gamma_t} = \frac{\dot{Y}_t}{Y_t} - \frac{\dot{V}_t}{V_t} = g(\beta)h(M) \tag{5}$$

The time index t is not suppressed as index to Y_t and V_t, because these variables will be used below at times with other indices. For the moment, all variables except prices are time dependent. The variable γ stands for variable factor productivity. The cost of a relative expansion of output equals $\beta h(M)$. The expression $g(\beta)h(M)$ stands for the rise in the productivity of variable factors of production. Equations (4) and (5) can be combined to give:

$$\dot{V}_t = V_t(\beta - g(\beta))h(M) \tag{6}$$

Notice the differences with the model of Kamien and Schwartz in section 4.1 of chapter 2: capital does not appear in the objective function and expansion of output is constrained by costs. The same optimal control method of solving the model that was used by Kamnien and Schwartz can again be employed (see appendix). This yields with respect to the direction of investment:

$$g'(\beta) = 1 - \frac{\int_t^T e^{-\rho s} P Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} \tag{7}$$

The numerator in the second term of the right hand side of equation (7) is the present value of future output, and the denominator is the present value of future costs of variable inputs. Together this term is some sort of intertemporal variable factor productivity. The slope of $g(\beta)$ in the optimum is negative, given that the present value of future output is larger than the present value of future variable costs, i.e. given that the present value of the future cash flow is positive. The larger future revenues, relative to future costs, the higher intertemporal variable factor productivity, the higher the optimal β and the lower the optimal $g(\beta)$, hence the more the firm will invest in expansion. The lower intertemporal variable factor productivity, the more the firm will invest in rationalization. The optimum must be the value for which is marginal revenue of investment in expansion equals marginal revenue of investment in rationalization.

With respect to the amount of investment we find:

$$h'(M) = \frac{e^{-\rho t}}{\beta \int_t^{\infty} e^{-\rho s} P Y_s ds - (\beta - g(\beta)) \int_t^{\infty} e^{-\rho s} w V_s ds} \tag{8}$$

Investment is expanded until the marginal revenue of investment (which is $h'(M)$ times the denominator in the right hand side of equation (8) above) equals its marginal cost. The optimal amount of present investment rises as the present value of future output rises. The effect of a rise in present value of variable costs on the optimal amount of investment depends on the difference between β and $g(\beta)$. If in equilibrium $\beta > g(\beta)$, then an increase of investment *raises* total variable factor costs, because the rise in variable factor productivity cannot compensate for the rise in variable factor demand due to expansion. If in equilibrium $\beta < g(\beta)$, then an increase of investment *decreases* total variable factor costs, because the rise in variable factor productivity more than compensates for the rise in variable factor demand due to expansion. Therefore, if in equilibrium $\beta > g(\beta)$ (i.e. the equilibrium direction of progress is biased toward expansion), a rise in the price of the variable input would decrease the optimal amount of investment, and if $\beta < g(\beta)$ (i.e. progress is biased toward rationalization), it would increase the optimal amount of investment.

We trace the behaviour of the optimal path for β, taking the time derivatives of equation (7):

$$\frac{dg'(\beta)}{dt} = \frac{e^{-\rho t}(PY_t + (g'(\beta) - 1)wV_t)}{\int_t^T e^{-\rho s} w V_s ds} \tag{9}$$

To determine the long term equilibrium value of β, we equate equation (9) to zero:

$$\frac{dg'(\beta)}{dt} = 0 \quad \Leftrightarrow \quad PY_t + (g'(\beta) - 1)wV_t = 0 \quad \Leftrightarrow \quad g'(\beta) = 1 - \frac{PY_t}{wV_t} \tag{10}$$

When the derivative $g'(\beta)$ is constant, then β is also constant, and so must be the ratio of total revenue PY_t and variable costs wV_t. This ratio is only constant, given that the firm invests ($M > 0$), when production Y_t and variable factors V_t grow with the same growth rate:

$$\frac{\dot{Y}_t}{Y_t} = \frac{\dot{V}_t}{V_t} \quad \Leftrightarrow \quad \beta h(M) = (\beta - g(\beta))h(M) \quad \Leftrightarrow \quad g(\beta) = 0 \tag{11}$$

In equilibrium there will be no investment any more in productivity increases, only in capacity expansion. Proof of the stability of the above equilibrium solution is given in the appendix.

To see what happens to the budget for investment M_t, consider the time derivative of equation (8):

$$\frac{dh'(M)}{dt} = h'(M)\{h'(M)((PY_t - wV_t)\beta + wV_t g(\beta)) - \rho\} \tag{12}$$

This expression can be used to trace the development of the optimal investment budget:

$$\frac{dM}{dt}=\frac{1}{h''(M)}\frac{dh'(M)}{dt}=\frac{h'(M)}{h''(M)}\{h'(M)((PY_t-wV_t)\beta+wV_tg(\beta))-\rho\} \qquad (13)$$

Let us consider the development when β has reached its long run equilibrium value and $g(\beta)=0$. Since $h'(M)>0$ and $h''(M)<0$, the condition for an ever growing budget M in the long run is:

$$\frac{dM}{dt}>0 \;\Leftrightarrow\; h'(M)\beta(PY_t-wV_t)<\rho \;\Leftrightarrow\; \frac{e^{-\rho t}\beta(PY_t-wV_t)}{\beta\int_t^{\infty}e^{-\rho s}(PY_s-wV_s)ds}<\rho \qquad (14)$$

That this condition is always fulfilled, can be seen by considering the following inequality:

$$\frac{e^{-\rho t}\beta(PY_t-wV_t)}{\beta\int_t^{\infty}e^{-\rho s}(PY_s-wV_s)ds}<\frac{e^{-\rho t}\beta(PY_t-wV_t)}{\beta\int_t^{\infty}e^{-\rho s}(PY_t-wV_t)ds}=\rho \qquad (15)$$

The inequality holds because, since $g(\beta)=0$, PY_t-wV_t grows over time. Thus, in long run equilibrium, ever larger amounts are spent on capacity expansion, without improving variable factor productivity any further.

This completes the outline of the reinterpreted model of Kamien and Schwartz, describing the intertemporal decision problem of a firm, confronted with opportunities for technological change. Given the restrictive assumptions of stable prices and fixed costs of technical progress, a firm will invest progressively more in capacity expansion, both in relative and in absolute terms.

3.4 Technical change as a cumulative phenomenon

Both the result that in the long term it will be optimal to invest only in capacity expansion, not in raising variable factor productivity, and the result that the optimal investment budget is ever expanding, are in conflict with intuition about real world situations. They arise because we assume, following Kamien and Schwartz, an extreme type of technological opportunity, in combination with price rigidity. Expressions (4) and (6) say that a *relative* expansion of productive capacity and a *relative* rise in variable factor productivity have a fixed price. Technical change can be pursued indefinitely for the same price. Nonetheless, we assume that the firm is a small agent in large and stationary markets, such that a decrease in production costs or an increase in variable factor demands is not translated into lower product prices and higher input prices.

In the next sections we shall adjust the specification of the model, to take account of these two features, such that continued technical advance gets progressively more expensive and such that prices react to changes in supply and demand. In this section, however, we shall retain the model as it has been presented as a bench-mark, and use it to clarify the cumulative character of technical change and to contrast technical progress with a situation of stationary technology.

3.4.1 A model without technical change

In the model of section 3 the firm faces opportunities for improvements in technology into the indefinite future. In this section, a model will be specified in which there is no technical progress, in which investments have no cumulative effect. In this model, the firm's moves are restricted to plain expansion and plain substitution, as defined in section 2. The actual possibilities of a firm in a real world situation must be somewhere in the middle: there is always some scope for introduction of improved technology but some investment will just go into doing more of the same. Therefore management, in drawing up their investment plan for the future, will consider these extreme models as bench-mark cases. Let us consider the decision problem for a firm, assuming no technical progress whatsoever.

$$\max_{\beta,M} Z = \int_0^T e^{-\rho t}(PY_t - wV_t - M)dt \tag{16}$$

subject to:

$$\dot{Y}_t = Y_0\beta h(M) \tag{17}$$

$$\dot{V}_t = \dot{Y}_t\frac{V_0}{Y_0} - V_0 g(\beta)h(M) = V_0(\beta - g(\beta))h(M) \tag{18}$$

The constants Y_0 and V_0 take the place of the variables Y_t and V_t. Equation (17) says no more than that an expansion of output always has the same price. Equation (18) says that demand for the variable factor of production changes by a constant times $(\beta - g(\beta))h(M)$. Variable factor requirements grow because of expansion and they additionally decline due to substitution. The model can be solved along the same lines again (compare equations (7) and (8)), yielding:

$$g'(\beta) = 1 - \frac{PY_0}{wV_0} \tag{19}$$

$$h'(M) = \left(\frac{\rho}{1 - e^{-\rho(T-t)}}\right)\frac{1}{\beta PY_0 - (\beta - g(\beta))wV_0} \tag{20}$$

Assuming the time horizon T to be distant, then $g'(\beta)$ and $h'(M)$ are independent of time. The investment budget is fixed and the ratio of expansion and substitution is constant over time. Both depend on prices and constants only.

We can compare the two situations described above, the occurrence of technical change and absence of technical change. First we will compare both models with respect to the ratio of capacity expansion and change in variable factor requirements, then with respect to the budget. We have seen that:

$$g'(\beta_m) = 1 - \frac{\int_t^T e^{-\rho s} P Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} \qquad \text{and} \qquad g'(\beta_z) = 1 - \frac{PY_0}{wV_0} \tag{21}$$

Here the index m stands for with technical progress and z stands for without. By comparing these two expressions it can be seen that $g'(\beta)$ with technical change is always smaller than without technical change, because the integral in the denominator will at most grow just as fast as the integral in the numerator. This is the case when β is at its maximum and $g(\beta)$ is zero. Starting at $t = 0$, however, it is expected that for some time $g(\beta)$ will be positive, and thus PY_t will grow faster than wV_t. Since $g'(\beta) < 0$ and also $g''(\beta) < 0$, it follows that without technical change β will always be smaller and $g(\beta)$ be larger than with technical change.

Now consider the research budget:

$$h'(M_m) = \frac{e^{-\rho t}}{\beta_m \int_t^T e^{-\rho s} P Y_s ds - (\beta_m - g(\beta_m)) \int_t^T e^{-\rho s} w V_s ds} \qquad \text{and} \tag{22}$$

$$h'(M_z) = \left(\frac{\rho}{1 - e^{-\rho(T-t)}} \right) \frac{1}{\beta_z P Y_0 - (\beta_z - g(\beta_z)) w V_0}$$

As time goes to infinity, the investment budget of the firm faced with technological change will rise indefinitely, as shown above. The investment budget of the firm without technical change is (nearly) constant, independent of time. Therefore, in the long run the investment budget with technical change will exceed the budget without. In the short term, under particular circumstances, the investment budget without technical change can exceed the budget with technical change. Since $h'(M) > 0$ and $h''(M) < 0$, it follows that as M grows $h'(M)$ gets smaller. The condition for a larger M without technical change at time t is:

$$h'(M_m) > h'(M_z) \quad \Leftrightarrow \tag{23}$$

$$\int_t^T e^{-\rho s}(\beta_m(PY_s - wV_s) + g(\beta_m) w V_s) ds < (\beta_z(PY_0 - wV_0) + g(\beta_z) w V_0) \int_t^T e^{-\rho s} ds$$

The inequality holds only if $g(\beta)$ and prices P and w would be such that e.g. for some time $\beta = 0$. The fact that it holds under these conditions is due to the rigid specification of the possibilities for substitution: the firm can substitute fixed factors indefinitely for variable factors, in exactly the same way and for the same price. In section 5 we shall come back to this problem. In general, however, the investment budget will be higher if there is technical progress, and in this section it is assumed that this is permanently the case.

3.4.2 Technological expectations

In this section, the two models presented above will be used, one describing a situation with technical progress and the other with a stable technology, to investigate the planning process of the firm and its outcomes. The models describe the planning process of the management of a firm at time zero. The management decides on a path for the investment budget and on a path for β and $g(\beta)$, the mixture of expanding capacity and reducing variable inputs per unit of output. The firm calculates its investment program and starts it at time zero. Suppose that, since changing firm strategies and investment plans is a costly matter, the plan is recalculated only at the beginning of time $t \gg 0$.

The management takes its decisions concerning β and M on the basis of its perception of the state of technology. If it is perceived that the firm is operating at the beginning of a technological trajectory, then it is expected that investment can result in productivity rises, in lasting effects through learning. Investment is supposed to have a cumulative effect. In this case the firm's planning problem is more closely described by the model *with* technical change. If, on the contrary, the management is pessimistic about the future of the present technological path, then they do not expect to go down any learning curve or experience any cumulative effects. They interpret the situation more alike the one described by the model *without* technical change.

The decisions of the management will be different, depending on how they judge the situation, more alike the one or more alike the other version of model. As the management sees possibilities for technical change, it will start to expand capacity at the expense of investing in variable factor saving. As technological outlooks are favourable, firms do not want to miss any of the new developments, because learning now is important for future performance. Variable factor saving, however, gets much more emphasis when the technological outlooks are dim. In that case the argument of learning does not apply. The firm tries to improve its margins by rationalizing, by replacing variable by fixed factors of production. The increasing adjustment costs, expressed by the concavity of $h(M)$, force the firm to proceed gradually, but by and by substitution takes place.

Expectations can be confirmed by experience or they can be disproved. In our context we can distinguish between four situations: technological change is expected or not, and the management's expectations turn out to be right or wrong. The situations can be grouped in a matrix. For each situation we can find an expression of what happens to output, variable factor requirements, variable factor productivity and profits. Let us look at output at time t, where t is such that the firm is still on its planned investment scheme initiated at time 0, and has not revised its plans in the light of new information. In that case, for the four possible combinations expectations and realizations concerning the occurrence of technical change, the result would be respectively:

Table 1: Output

	t.c.		no t.c.
t.c. expected	$Y_0 + \int_0^t Y_s \beta_s h(M_s) ds$	>	$Y_0 + Y_0 \int_0^t \beta_s h(M_s) ds$
	$\vee$		$\vee$
t.c. not expected	$Y_0 + \beta_z h(M_z) \int_0^t Y_s ds$	>	$Y_0 + Y_0 t \beta_z h(M_z)$

It can easily be seen that technological change always results in higher output than no technical change, whatever the expectations, since Y_t grows with technical change. Also, the expectation of technological change always results in higher output, whether there is in fact technological change or not, since both β and the budget *M* are higher when technological change is expected.

Next we look at variable factor requirements:

Table 2: Variable input requirements

	t.c.		no t.c.
t.c. expected	$V_0 + \int_0^t V_s(\beta_s - g(\beta_s))h(M_s)ds$	$>$	$V_0 + V_0 \int_0^t (\beta_s - g(\beta_s))h(M_s)ds$
	$\vee(\wedge)$		$\vee(\wedge)$
t.c. not expected	$V_0 + (\beta_z - g(\beta_z))h(M_z) \int_0^t V_s ds$	$>$	$V_0 + V_0 t(\beta_z - g(\beta_z))h(M_z)$

Technological change, whether expected or not, increases variable factor demand. The expectation of technological change, whether right or wrong, in general increases variable input demand also. Only if $\beta_s - g(\beta_s) < 0$ and $h(M_s) \gg h(M_z)$, the expectation of technological change could lead to lower variable factor demand. In other words, an expected progress in technology will induce firms to expand their investment budgets and shift their efforts to a larger expansion in output capacity, relative to economizing on variable inputs: $\beta/g(\beta)$ rises. If, however, on balance the firm would decrease variable input requirements in absolute terms, because $\beta < g(\beta)$, the higher investment budget when technical change is expected might over-compensate the higher $\beta/g(\beta)$ ratio. Given the assumptions about a stable environment, expressed in fixed prices, the message is clear: optimistic expectations about technical progress boasts the economy, and so does actual technical change. It stimulates production, variable factor demand and investment.

Now we turn to variable factor productivity Y_t/V_t (see Table 3). In general, the ordering of the numerator is the same as the ordering of the denominator. However, it is easy to see that if there is technological change, whether it is expected or not, then output will grow faster than variable factor requirements. Variable factor productivity increases if there is technical change. The influence of expectations, on the contrary, is not clearly determined. In case there is no technological change (the right column), the condition under which productivity will grow faster if technological change is not expected then when it is expected, is:

$$\frac{1 + \int_0^t \beta_s h(M_s)ds}{\int_0^t g(\beta_s)h(M_s)ds} > \frac{1 + t\beta_z h(M_z)}{tg(\beta_z)h(M_z)} \tag{24}$$

This condition holds, given that the budget is larger and β is larger when technical progress is expected. Determination of what happens when technological change does occur (the left column) is impossible in general terms. There are effects and counter effects and it cannot be determined in general what the outcome on balance will be. The situation can be such that the expectation of technical change leads to a higher variable factor productivity than if the technical change comes as a surprise. Then, overlooking the whole matrix, it can be concluded that the right expectation leads to higher variable factor productivity and wrong expectations depress variable factor productivity. However, the situation can also be reversed. The circumstances can be such that the expectation of technical change leads to a lower variable factor productivity than if technical change comes as a surprise, since the expectation gears investments toward expansion.

Table 3: Variable factor productivity

	t.c.		no t.c.
t.c. expected	$\dfrac{Y_0+\int_0^t Y_s\beta_s h(M_s)ds}{V_0+\int_0^t V_s(\beta_s-g(\beta_s))h(M_s)ds}$	$>$	$\dfrac{Y_0+Y_0\int_0^t \beta_s h(M_s)ds}{V_0+V_0\int_0^t(\beta_s-g(\beta_s))h(M_s)ds}$
	$\vee\wedge$		$\wedge$
t.c. not expected	$\dfrac{Y_0+\beta_z h(M_z)\int_0^t Y_s ds}{V_0+(\beta_z-g(\beta_z))h(M_z)\int_0^t V_s ds}$	$>$	$\dfrac{Y_0+Y_0t\beta_z h(M_z)}{V_0+V_0t(\beta_z-g(\beta_z))h(M_z)}$

Thus we arrive at the conclusion that the expectation of technical change, irrespective of whether it is justified or not, could depress variable factor productivity. The prospect of technological advance makes the firm willing to accept a more wasteful use of variable factors like energy, raw materials, intermediaries and labour, than otherwise, in exchange for the chance to get acquainted with new production techniques.

Finally, profits under the four different regimes could be ordered. It then turns out that, profits are higher with technical change than without, and they are higher with the right expectations than with the wrong ones.

In this section two models representing bench-mark cases were compared. The one model portrays firm planning in times of relentless technological progress, the other model describes firm planning in times of stationary technology. The models were then used to explore the effects of technological expectations on production, variable factor demand and variable factor productivity. The model shows that the expectation of technical progress can depress variable factor productivity, first of all when technological progress is rightly expected, but even more so when it is wrongly expected.

The above approach might help to interpret the investment policy of firms in sectors with irregularly developing technology. An example might be found among firms in micro-electronics. In the semi-conductor industry there has been a shake out of producers and a wave of reorganizations taking place in the recent past. It all started with high technological expectations: chip technology was supposed to move very fast. This could be an argument to invest rather massively in this technology. As in our model, it was assumed that the benefits of the investment are in the learning effects, not in the immediate results: investment puts one in a better position for the future. So, given technological expectations it is thoroughly rational to expand capacity fast, although this could imply a relatively larger expansion of the variable factor requirements and less economizing on their use. This keeps productivity below the level that would have been attained when expectations were more moderate. In case the technological expectations happen to be unjustified, the firm ends up in the least favourable corner of the matrix. Both variable factor productivity and profits go down to the minimum values. Exaggerated expectations might thus easily lead to falls in profits. The situation might even be aggravated, when price effects would be included in the model. If many firms have optimistic expectations and expand rapidly, overcapacity is bound to put prices under pressure, depleting profit margins further.

3.5 Depletion of technological opportunities

The firm model presented in section 3 suffers from its rather strong assumptions on the persistence of technical progress. This leads to unlikely conclusions about developments in the long term: ever larger amounts are invested, exclusively in capacity expansion. In this section, the constraints describing technological progress will be modified, such that a decrease in the perceived opportunities for technical progress, and finally even a limit to further exploitation of a technological trajectory, are incorporated. The specification of the constraints as they stand, is such that there are decreasing returns to scale of investments in technical change *in* time, at some period *t*, but increasing returns to scale *over* time. This explains why it is most profitable in the long term to expand capacity only. The larger the gap between total revenues and total variable costs, the more profitable it is to increase capacity at the expense of less increasing variable factor productivity. But, the more expansion, the larger the gap. After a while the gap between total revenue and variable costs gets so large, that a percentage expansion of capacity is always more profitable than raising productivity, whatever the costs.

The assumption of a constant price for a relative improvement is in conflict with common experience, as expressed in Wolff's Law. As Freeman notes: 'Wolff was a German economist who in 1912 published four "laws of retardation of progress". Essentially, he argued that the scope for improvement in any technology is limited, and that the cost of incremental improvement increases as technology approaches its long run performance level.' [Freeman 1980, p.216 note 2]. In general, technical improvements get progressively more expensive, as easy solutions start to be exhausted. One can in this context recall the concept of a technological trajectory. As a firm proceeds along a technological path, based on a specific basic innovation, it gradually depletes the possibilities for further improvement. Only when a new basic innovation is launched, new profitable possibilities for advancement arise.

All this does not imply that technical change on the aggregate level also gets progressively more expensive. In the economic system as a whole, new technological trajectories open up at irregular and unpredictable intervals. The point is here that the appearance of any specific basic innovations, necessary for the firm, is highly uncertain and cannot be planned by the firm. An analysis of the problem is also provided by Evenson and Kislev [referred to by Binswanger, 1978, pp. 92-97], who describe research as a sampling process. Magat (1979) tries to improve upon the Kamien and Schwartz model, arguing that '[..] the current values of the augmentation parameters, A and B, summarise the past R&D expenditure pattern, and thus the position of the frontier should depend upon their values.' The innovation-possibilities frontier is then made dependent on factor augmentation parameters A and B in the following way (compare equation (4) in chapter 2):

$$\frac{\dot{A}}{A} = g(\beta)u(A) \qquad \frac{\dot{B}}{B} = \beta v(B) \tag{25}$$

Allowing for exhaustion of innovative opportunities can introduce a bias in the direction of technical change. Magat's adjustments of the Kamien and Schwartz model are not completely satisfactory, because they do not allow for an analytical solution of the model. In this section I shall attempt to introduce an approximate expression of Wolff's Law in the model, which does allow us to solve the model.

The age and the technological level of the present capital goods, as well as the history of expenses on R&D are given for the firm at the time of decision making. They enter the formulation of the decision problem of the firm through the constraints. The technological level inherited from the past determines the possibilities in the future. A way to insert this into the model would be to make the effectiveness of investment dependent upon the history of investments. We expand the model as follows:

$$\underset{\beta,M}{\text{Max}}\, Z = \int_0^T e^{-\rho t}(P\,Y_t - w\,V_t - M)dt \tag{26}$$

subject to:

$$\dot{Y}_t = Y_t\beta h(M,\tau) \tag{27}$$

$$\dot{V}_t = V_t(\beta - g(\beta))h(M,\tau) \tag{28}$$

$$\dot{\tau}_t = h(M,\tau) \tag{29}$$

Assume that the firm proceeds on a technological trajectory, and progress becomes more expensive as the limits of the trajectory are nearing. The variable τ can then be an index for the 'technological level' of the firm, relative to its trajectory. As τ rises, the distance of the firm to the end of its current trajectory decreases. Thus the present 'technical position on the current trajectory' τ determines the effectiveness of investments in technical progress. A switch from an old technological trajectory to a new one would be expressed by a quantum jump of τ downward. The index for the technical level itself must grow as the firm invests more in technical advance. A measure for the advance in a certain period is the effectiveness of investment in that period, expressed by the function $h(M,\tau)$. Therefore this function is assumed to determine the change in technical level.

An important question is, how exactly the technical index τ determines the speed of technical progress $h(M,\tau)$. It can be argued, that at first there might be some increasing returns to investment in innovation along a new technological path, but the longer innovation along the same technological trajectory proceeds, the less effect investments have on productivity and capacity. Thus the second derivative of $h(M,\tau)$ with respect to τ must be negative. This condition ensures that $h(M,\tau)$ will decrease and approach zero as τ increases, indicating the end of possible progress along the current trajectory. When the model is being solved (see appendix), it turns out that $g'(\beta)$ remains the same as in section 3, indicating that with this specification the direction of technical progress is not influenced by depletion of technical opportunities. The derivative of $h(M,\tau)$ with respect to M turns out to be:

$$h_M = \frac{e^{-\rho t}}{\int_t^T e^{-\rho s}\left\{\beta P Y_s - (\beta - g(\beta)) w V_s + \frac{h_\tau}{h_M}\right\} ds} \tag{30}$$

The change of this derivative in time is:

$$\dot{h}_M = h_M\{h_M(\beta P Y_t - (\beta - g(\beta)) w V_t) + h_\tau - \rho\} \tag{31}$$

From this expression we can try to trace the development of the optimal investment budget:

$$\dot{h}_M = h_{MM}\dot{M} + h_{M\tau}\dot{\tau} \tag{32}$$

The change of the slope of $h(M,\tau)$ over time is attributable to a change in the optimal budget M and to a change in the position of the function h due to increasing technical level τ. The optimal budget declines if:

$$\dot{M} = \frac{1}{h_{MM}}(\dot{h}_M - h_{M\tau}\dot{\tau}) < 0 \quad \Leftrightarrow \quad \dot{h}_M > h_{M\tau}\dot{\tau} \tag{33}$$

Substitution gives:

$$h_M\{h_M(\beta P Y_t - (\beta - g(\beta)) w V_t) + h_\tau - \rho\} > h_{M\tau}\dot{\tau} \tag{34}$$

Now assume, for mathematical convenience, that $h(M,\tau)$ is a separable function: $h(M,\tau) = h_1(M)h_2(\tau)$. In that case we can write:

$$h_M h_\tau = h_{M\tau} h = h_{M\tau}\dot{\tau} \tag{35}$$

Substitution of equation (35) reduces (34) to:

$$h_M(\beta P Y_t - (\beta - g(\beta)) w V_t) > \rho \tag{36}$$

Like in the previous section we can substitute for h_M and evaluate the inequality at $g(\beta) = 0$:

$$\frac{e^{-\rho t}\beta(P Y_t - w V_t)}{\beta \int_t^T e^{-\rho s}\left(P Y_s - w V_s + \frac{h_{\tau_s}}{h_{M_s}}\right) ds} > \rho \tag{37}$$

As the firm approaches the technological frontier, h_τ is negative, going to minus infinity, and h_M is positive. $PY_t - wV_t$ is finite in finite time (and might only go to infinity as time goes to infinity). Since h_τ goes to minus infinity in finite time, as τ approaches its border value, and h_M does not go to plus infinity before M equals zero, the denominator on the left hand side approaches zero. Thus from some moment in time onward the inequality holds.

The speed of technical progress is measured by $h(M,\tau)$, which develops in time as:

$$\dot{h}(M,\tau) = h_M\dot{M} + h_\tau\dot{\tau} \tag{38}$$

Given that h_M and $\dot{\tau}$ are positive and, after a certain period of technical progress, both h_τ and $\dot{M}$ are negative, technical progress slows down in the long run.

In conclusion we can state that a firm will proceed along a technological path with variable speed. At first M goes up as returns to innovative investments increase. The speed $h(M,\tau)$ increases. Then the effectiveness of these investments declines more and more. As $h(M,\tau)$ declines because a higher technical level is reached, this induces the firm to decrease M, such that the speed of technical progress $h(M,\tau)$ decreases even faster. The speed of innovation first goes up slowly but eventually runs down quickly.

The model without technical progress, presented in section 4.2, suffers from a similar weakness as the model with technical change. In that model it is assumed that substitution of variable inputs by fixed investments can go on forever in the same way: the total requirement of variable inputs is made up of the same units that can be replaced by the same units of fixed investments, say every worker can be replaced by a similar machine. Obviously, the last worker cannot be substituted like the first, since he has to control the machinery. A similar amendment proposed for the model with technical change can be used in the model with stationary technology too, introducing depletion of opportunities for substitution and replication.

Exhaustion of innovative opportunities for a firm operating on a specific technological trajectory is one factor that is likely to influence the effectiveness of an amount of investment M. Another factor, working in the opposite direction, is the possibility of technology spill-overs from firms that are more advanced, further along the technological path. Firms can learn from their more efficient competitors. This can be included in the model in a straightforward way: suppose the technological level of the firm at the present technological frontier is θ, then the function expressing the effectiveness of investment M, for a firm of technological level τ, in an industry where the highest level is θ, would be $h(M,\tau,\theta)$, where θ changes in the competitive struggle.

3.6 Endogenously determined prices

In section 5 the basic model was extended to allow for depletion of technological opportunities. In this section another restrictive assumption will be relaxed: instead of assuming constant prices, we shall assume that they are determined by supply and demand. Gradually the model gains in realism, which will allow some further inferences about the firm's decision making. By assuming prices to respond to the firm's actions, an element of competition on the market is introduced: the firm does not operate in a void any more, since the production volume that competitors put on the market influences the prices the firm can get for its products. For purposes of exposition, it is assumed in the first part of this section that only the output price is flexible. The consequences of price flexibility are then discussed on an intuitive basis. In the second part, rivalry is also introduced on the market for variable inputs and the preceding intuitive analysis is complemented by a more rigourous treatment.

3.6.1 A downward sloping demand curve

The assumption that the firm operates in a competitive environment, where the demand side has an infinite capacity to absorb the supply of output against the fixed price P, is now be replaced by a more general representation of demand: it is assumed that the firm takes demand to follow a demand curve characterized by a constant price elasticity. The inverse of this price elasticity of demand is indicated by the parameter $-\pi$. Thus an increase of 1% in industry output causes a π% reduction in price. The production of output produced by other firms than the firm which we model is Q_t, such that total production at time t equals $Q_t + Y_t$. The model can now be reformulated as follows:

$$\underset{\beta,M}{\text{Max}}\, Z = \int_0^T e^{-\rho t}(P_t Y_t - w V_t - M)dt \tag{39}$$

subject to:

$$\dot{Y}_t = Y_t \beta h(M) \tag{40}$$

$$\dot{V}_t = V_t(\beta - g(\beta))h(M) \tag{41}$$

$$\frac{\dot{P}_t}{P_t} = -\pi \frac{\dot{Q}_t + \dot{Y}_t}{Q_t + Y_t} \quad \Leftrightarrow \quad \dot{P}_t = -\pi P_t \frac{\dot{Q}_t + Y_t \beta h(M)}{Q_t + Y_t} \tag{42}$$

Solving this problem (see Appendix) yields the following equilibrium conditions for the changes of the slope of $g(\beta)$ and $h(M)$ over time:

$$g'(\beta) = 1 - \frac{\int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} \tag{43}$$

$$h'(M) = \frac{e^{-\rho t}}{\beta \int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds - (\beta - g(\beta)) \int_t^T e^{-\rho s} w V_s ds} \tag{44}$$

These expressions differ only marginally from our results in section 3, equations (7) and (8): in both expressions, $\left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right)$, which is 1 minus π times the firm's market share at time s, serves as a weight of $e^{-\rho s} P_s Y_s$. The fact that the demand curve is downward sloping modifies the weight of future revenues in equations (7) and (8), but not the weight of future variable costs. It can readily be seen that this model reduces to the fixed output price model of section 3, if either the inverse price elasticity π is zero, or the market share of the firm goes to zero, such that the firm has no substantial influence on the equilibrium product price.

In equations (43) and (44), which determine the planned paths for the direction of development $g(\beta)/\beta$ and the investment budget M, the future output of the rest of the industry Q_t, appears as an argument. Thus the firm has to formulate expectations on the future output growth of competitors in the industry. Here two types of assumptions are possible. Expectations can either be based on past or current observables like growth rates, or they can be based on a model of strategic interaction (say rational expectations). The consequences of assuming one or the other mechanism will not be explored here: it is merely supposed that the firm *has* expectations, not that they are correct, nor that they are rational, nor that they are consistent with competitors' expectations; it is only assumed that the firm thinks that competitors future plans are not influenced by changes in its own plans.[11] Nevertheless, the notion of limited information and binding cognitive constraints would suggest that an expectations formation mechanisms of the first type would be a more promising hypothesis if one aims at explaining empirical patterns of competition.[12] To see what conclusions can be drawn from this model concerning innovative activity and competition, we analyse the last equations from two different angles. First we assume a *firm* of a given size, and examine how its behaviour would change as the size of the market would change; then we assume a *market* of a given size and examine how firm behaviour would differ as this market is split up among more or less firms.

First of all, consider what a firm of a specific size would do in a market of different sizes. Suppose that output volume Y_0 of the firm at time 0 is given. For larger values of Q_t in equation (44), which is equivalent to a smaller market share of the firm, $h'(M)$ gets smaller, which means that since $h(M)$ is concave M grows. Thus the larger the aggregate production by competitors, the more a firm will invests. Moreover, the investment will be directed more toward expansion than toward rationalization: a larger Q_t leads to a smaller (further negative) $g'(\beta)$, according to equation (43), which means that β gets larger, since $g(\beta)$ is concave. This result is plausible, for it says that when the firm is smaller relative to the total market, it has less influence on the output price, and thus suffers less from the fact that growth of output leads to a fall in price. Therefore it will expand more and put relatively less effort in reducing variable factor costs.

Alternatively, suppose there is a market of a certain size $Q_0 + Y_0$ at time $t = 0$, in which a number of n identical firms compete, and in which there is no exit nor threat of entry. Under these conditions, the firm we model is a representative firm. Indicate the total market by $n\,Y_t = \overline{Q}_t$. Each firm requires V_t variable inputs and the total requirement at time t is $n\,V_t = \overline{X}_t$. The market share of each firm at all times t equals $1/n$. Equations (43) and (44) can thus be written as follows:

[11] Like in Cournot competition, the firm treats its competitors' output supply and variable input demand decisions for periods 0 to T as given, and plans an optimal response. Bertrand competition is not explored here, because it would not fit to the assumption that present volume decisions are largely dependent on past volumes.

[12] Assuming that firms are able to form rational expectations in the context of this model would mean that firms are supposed to be familiar with all current actions and future technological options of all their competitors, and also with their expectations concerning volumes, prices and interest rates.

$$g'(\beta) = 1 - \frac{\int_t^T \left(1 - \frac{\pi}{n}\right) e^{-\rho s} P_s \frac{\overline{Q}_s}{n} ds}{\int_t^T e^{-\rho s} w \frac{\overline{X}_s}{n} ds} = 1 - \frac{\left(1 - \frac{\pi}{n}\right) \int_t^T e^{-\rho s} P_s \overline{Q}_s ds}{\int_t^T e^{-\rho s} w \overline{X}_s ds} \tag{45}$$

$$h'(M) = \frac{e^{-\rho t}}{\beta \int_t^T \left(1 - \frac{\pi}{n}\right) e^{-\rho s} P_s \frac{\overline{Q}_s}{n} ds - (\beta - g(\beta)) \int_t^T e^{-\rho s} w \frac{\overline{X}_s}{n} ds} = \tag{46}$$

$$\frac{n e^{-\rho t}}{\beta\left(1 - \frac{\pi}{n}\right) \int_t^T e^{-\rho s} P_s \overline{Q}_s ds - (\beta - g(\beta)) \int_t^T e^{-\rho s} w \overline{X}_s ds}$$

From (46) it is clear that it depends on the price elasticity of demand $1/\pi$ whether investment is larger and whether output grows faster in a more or in a less concentrated market. If this elasticity is ***high*** (the demand curve is flat), and thus π is close to zero in equation (46), then a large number of firms causes each of them to invest a lower budget M than a monopolist would do. Thus each of their output volumes would grow by a lower percentage than the output of the monopolist would grow. Therefore the total of output would grow less than under monopoly. The monopolist invests *more* because he is larger than a firm with only a share of the total market (of which the size is given), and therefore enjoys economies of scale in investment: a large firm can reap more benefits from its investment than a smaller firm. The extent to which smaller firms would invest less than the monopolist would do depends on the curvature of $h(M)$. The more concave $h(M)$, the less the difference between the investments of the monopolist and the firm with only part of the market.

If, however, the elasticity is *low* and π is not close to zero (a steep demand curve), the opposite might come about. Then, as the denominator in equation (46) gets very small and approaches zero, a monopolist or oligopolist might decide to invest not at all in expansion of output. If there is any investment, it will be directed as much as possible toward variable input saving, as is apparent from (45). Under these conditions the industry might grow faster, if there are more rather than less competitors in the market. Here the monopolist invests *less* because expansion of output would result in a substantial decrease of the output price, which would undermine the returns it earns at present scale of its operations. If π is large enough, if demand is sufficiently inelastic, then for a low number of firms n (provided that $\beta - g(\beta)$ is no too small) $h'(M)$ will become non-positive and M thus zero, indicating that it may be optimal not to invest at all.

3.6.2 Variable factor prices

Similarly, the model can be extended by assuming oligopolistic competition on factor markets. Suppose that the supply curve for variable inputs is characterized by a constant price elasticity $1/\phi$. The variable factor requirements by all the other producers in the industry are indicated by X_t. The model is now:

$$\underset{\beta,M}{\text{Max}}\, Z = \int_0^T e^{-\rho t}(P_t Y_t - w_t V_t - M) dt \tag{47}$$

subject to:

$$\dot{Y}_t = Y_t\beta h(M) \tag{48}$$

$$\dot{V}_t = V_t(\beta - g(\beta))h(M) \tag{49}$$

$$\dot{P}_t = -\pi P_t \frac{\dot{Q}_t + Y_t\beta h(M)}{Q_t + Y_t} \tag{50}$$

$$\dot{w}_t = \phi w_t \frac{\dot{X}_t + V_t(\beta - g(\beta))h(M)}{X_t + V_t} \tag{51}$$

The model can be solved, yielding:

$$g'(\beta) = 1 - \frac{\int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds}{\int_t^T \left(1 + \phi \frac{V_s}{X_s + V_s}\right) e^{-\rho s} w_s V_s ds} \tag{52}$$

$$h'(M_t) = \frac{e^{-\rho t}}{\beta \int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds - (\beta - g(\beta)) \int_t^T \left(1 + \phi \frac{V_s}{X_s + V_s}\right) e^{-\rho s} w_s V_s ds} \tag{53}$$

The analysis in the previous section already clarified the most important consequences of assuming variable prices. Below the conclusions will be supplemented and elaborated in five steps. We shall focus on the links between *elasticities* and investment planning, on *expectations* as determinants of investment, on the role of *firm size* and *market concentration* and the Schumpeterian hypotheses, and on the conditions for *steady state growth*.

First of all, consider the effect of variation in the *elasticity parameters* π and ϕ on a firm's investment plans. A more price inelastic supply of variable factors (ϕ large), other things equal, has a similar effect on the direction of technical change as a price inelastic demand for output (π large): investment will be more directed toward cost saving ($g'(\beta)$ will be closer to zero and $g(\beta)$ will be large; see equation (52)). If the optimal direction of technical change will be such that $\beta < g(\beta)$, then a larger ϕ will lead to a larger investment budget M. On the contrary, if $\beta > g(\beta)$, then a larger ϕ leads to a smaller budget M.

Intuitively these results are plausible: inelastic demand and supply curves mean that changes in input and output quantities have large effects on prices. If these effects are large, then it is best not to expand output too much, in order to keep the output price high, and to economize on inputs, in order to keep the input price low. If technical possibilities are such that it is relatively easy and cheap to cut back variable input requirements per unit of output, leading to a decrease in variable factor requirements in absolute terms (which depends on the shape of $g(\beta)$) then firms will expand their budget M. If, however, $g(\beta)$ has such a curvature that it remains optimal to expand variable factor requirements in absolute terms, then firms will decrease their investment budgets M.

Secondly, let us look at the relationship between *market concentration* and investment behaviour. Consider once more a market of given size $\overline{Q}_0$ at time 0, equally partitioned among *n* identical firms. Each firm uses the same amount of inputs, which sum up to $\overline{X}_0$ at time 0. Given *n* equal firms, equations (52) and (53) can be written as:

$$g'(\beta) = 1 - \frac{\left(1 - \frac{\pi}{n}\right)\int_t^T e^{-\rho s} P_s \overline{Q}_s ds}{\left(1 + \frac{\phi}{n}\right)\int_t^T e^{-\rho s} w_s \overline{X}_s ds} \tag{54}$$

$$h'(M) = \frac{n e^{-\rho t}}{\beta\left(1 - \frac{\pi}{n}\right)\int_t^T e^{-\rho s} P_s \overline{Q}_s ds - (\beta - g(\beta))\left(1 + \frac{\phi}{n}\right)\int_t^T e^{-\rho s} w_s \overline{X}_s ds} \tag{55}$$

From equations (54) and (55) it is clear that a smaller number of firms *n* has the same consequences as lower price elasticities: the effects of relative input and output changes on prices have a larger impact on the individual firm. Thus firms put more effort in lowering input demands and less in expanding output. It can be seen from equation (55), that if π and ϕ are close to zero, the *more* firms are in the market, the *lower* the investment budget of each competitor is. However, if π is large, then an increase in the number of firms *n* may lead to *higher* investments by *each* of them. Furthermore, it can be shown using equation (54) that if output demand is inelastic (π large), but input supply is elastic (ϕ small), an *increase* in the number of competitors *n* in the market may lead each of them to redirect investment toward expansion (see Appendix).

Thirdly, interpret the model as a representation of *investment planning* of a firm in a competitive market, where not all firms are of equal size at time 0 and where entry or exit might occur. The expected actions of competitors enter the determination of the equilibrium β and *M* of the firm through the expected future prices P_s and w_s and through the expected development of its market shares $Y_s/(Q_s + Y_s)$ and $V_s/(X_s + V_s)$. Suppose a firm expects the output of the rest of the industry Q_s to grow by a rate ψ and the variable input demand of competitors to grow by a rate ξ. Variation in these values can be related to the threat of entry or to expectations with respect to competitors' successes in innovation. The firm thus *expects* $Q_t = Q_0 e^{\psi t}$ and $X_t = X_0 e^{\xi t}$, where Q_0 and X_0 are *observed* competing output supply and input demand at the moment it calculates its investment plan.

It can be shown that the reaction of a firm to a perceived higher growth rate of competing output supply depends on the elasticity parameter π and the firm's market share. On the one hand, if the firm expects to have a *large* share of the market and if it assumes that output demand is inelastic, then the expectation of a *faster growth of competing supply* will induce the firm to *increase* its investments in expansion. The firm in this situation will engage in 'competitive warfare', because the 'advantage' of a foreseen loss of market share outweighs the loss due to a drop in prices. A loss of market share is an advantage here because it means that a price decrease is less internalized; as the firm expects to have less market share, it considers the price drop to a larger extent an externality. Formally, if for all periods *s*, where $t \leq s \leq T$, the firm expects $Y_s > Q_s/\pi$ (its own output larger than competing output times the price elasticity of demand), then $dM_t/d\psi > 0$ (see Appendix). Similarly, if the firm expects competitors to decrease their growth rate, it will cut down its own rate too. On the other hand, if the firm expects to have a *small* market share and if it assumes that demand is elastic, then the expectation of a *faster growth of competing supply* will induce the firm *to cut expansionary investment*. The firm now backs down, calculating that the price drop justifies less expansion, while the loss of market share does not induce the firm much to increase expansion.

On the input side matters are less ambiguous: a larger expected increase of variable input demand induces a firm to increase investment in rationalization and to cut down expenses on expansion (see Appendix). As a larger input demand is expected, the firm expects both rising prices and a fall in its share on the market for inputs. The inducement to step up investment spending in rationalization upon an expected input price increase turns out always to outweigh the stimulus to cut down investments because of a smaller market share. Finally, it may be noted that the above conclusions do not depend on the cumulative effects of investment and therefore hold in a model *with* as well as *without* technological opportunities (i.e. a model along the lines of section 4.1, assuming variable prices though).

Fourthly, let us try to use the model to assess the issue of the relationship between the speed of technological progress, *market structure* and *firm size*, and thereby shed some light on the so-called Schumpeterian hypotheses [see e.g. Kamien and Schwartz (1982)]. These postulate a positive relationship between the speed of technological progress and both market concentration and firm size. In the above model, there are two forces which exert an influence on the speed of technological progress: economies of scale and the degree of internalization of price decreases on input and output markets. Economies of scale depend on firm size and the degree of internalization of a price decrease depend on market share, and thus market structure. The direction of these two influences depends on the *type* of technological change. We argued that technological progress in a firm can take two forms: improvement of the capital stock, which leads to a higher production capacity, and improvement of the use of variable inputs, which leads to a reduction in variable factor demand per unit of output.

Consider first the *former* type of technological progress, improvement of capital leading to extra capacity. On the one hand, a large firm in a concentrated market takes more advantage of economies of scale than a small firm in a market of the same size, and will therefore invest more in capacity expanding technological progress. On the other hand, a firm with a large market share internalizes the negative influence of a price decrease on industry revenues to a larger extent, and therefore restricts investment in capacity expanding technological progress. This is expressed by the *positive* sign of Y_t in equations (52) and (53), and the *negative* sign of $Y_t/(Q_t + Y_t)$. The influence of these two *firm specific* factors is moderated by two factors which are *common* to the market: size Y_t is multiplied by output price P_t and market share $Y_t/(Q_t + Y_t)$ is multiplied by elasticity parameter π. The higher the output price and the elasticity of demand, the more attractive it is to invest in extra capacity for *all* firms in the market.

Equations (52) and (53) thus show that the dynamics of investment are related to changes over time in *four* variables on the output side. As a market develops and the firm invests, the firm's output Y_t rises. What happens to its market share depends on other firms' behaviour. As *aggregate* output expands, the price P_t falls (given a stable demand curve) and demand gets less price elastic: π rises. As prices fall, the firm size effect on investment, inducing the firm to invest more as it grows, loses strength. Simultaneously, due to a rise in π, the market share effect, inducing the firm to invest less as it has a larger market share, rises in strength. Firms are therefore likely over the course of the development of a market or an industry first to raise investment in expansion, reaping the advantages of scale economies related to their own growth, and then to decrease investment, as the effects of a price squeeze get more severe.

Next consider the *latter* type of technological change, improvement of variable factor productivity. Here both tendencies work in the same direction. A large firm (in terms of V_i) invests more in rationalization, because it can reap more benefits due to economies of scale. In addition, a large firm, if it exerts a large share of demand $V_i/(X_i + V_i)$ in the variable factor market, also invests more, because it fully internalizes the beneficial effect of a drop in the unit costs of variable factors. Thus in investments in rationalization large firms in concentrated markets may progress fastest. Again the two factors are moderated by the relevant price w_i and elasticity parameter ϕ, which are likely to change as the industry matures.

It follows from the present model that absolute size is positively related to the speed of technological progress, but relative size (market share) negatively. This disagrees with the Schumpeterian hypotheses mentioned before, that claim a positive relationship in both instances. There has been a fair deal of empirical research on the validity of the Schumpeterian hypotheses, and the evidence is rather mixed, as noted in section 3.2 of chapter 2 [Scherer (1980), Cohen and Levin (1989)]. Reasons for the ambiguous and inconclusive nature of the empirical evidence on these hypotheses which can be suggested on the basis of the present model are: 1) the fact that there are good arguments for the conjecture that the second Schumpeterian hypothesis is (partly) wrong (as far as investment in progress generates extra output which pushes down the price); 2) the fact that size and market share influence investment simultaneously and in opposing directions, but are mostly severely correlated, hampering empirical testing; and 3) the fact that the relative strength of the 'size factor' and the 'market structure factor' varies over the life-cycle of the industry.[13]

Finally, consider the issue of a long term *steady state growth* path. Our result in section 3, assuming fixed prices on input and output markets, was that in the long term there is no steady state growth: $g(\beta)$ goes to zero, the budget M rises indefinitely and output grows explosively. If prices are flexible, though, we saw that investments budgets may decrease over time and go (asymptotically) to zero, as firms internalize the negative effects on revenues of an output price decrease. Investment decreases if demand is inelastic, if π is large. Demand tends to get inelastic, as time progresses, markets expand and get saturated. When investment ceases, the technological development comes to a standstill and capital moves to other investment opportunities. This is a familiar situation in many markets.

[13] One might object that Schumpeter, writing on innovation and technical change, had foremost qualitative progress in output, product innovations, in mind [see e.g. Schumpeter (1943), chapter 7], where the model assumes a homogeneous output. The model may be reinterpreted, however, regarding output as the fulfilling of a need (e.g. information, music, transport, etc.) by means of a product which can be improved over time.

Let us consider, however, the more simple case in which elasticity parameters π and ϕ remain constant over time, and see whether a steady state growth path of the firm can be determined.[14] Call the steady state growth rate of output $\dot{Y}_t/Y_t = \mu$ and the steady state growth rate of variable input demand $\dot{V}_t/V_t = \nu$, the steady state growth of productivity being $\mu - \nu$. Thus for output and variable inputs we have: $Y_t = Y_0e^{\mu t}$ and $V_t = V_0e^{\nu t}$. In steady state, market shares are constant and therefore competing output Q_t and variable factor demand X_t also grow with rates μ and ν. From equations (50) and (51) it follows that prices in steady state are: $P_t = P_0e^{-\pi\mu t}$ and $w_t = w_0e^{\phi\nu t}$. Substituting these expressions in equations (52) and (53) we get:

$$g'(\beta) = 1 - \frac{\left(1 - \pi\frac{Y_t}{Q_t + Y_t}\right)P_0Y_0\int_t^T e^{((1-\pi)\mu-\rho)s}ds}{\left(1 + \phi\frac{V_t}{X_t + V_t}\right)w_0V_0\int_t^T e^{((1+\phi)\nu-\rho)s}ds} \tag{56}$$

$$h'(M_t) = \frac{e^{-\rho t}}{\beta\left(1 - \pi\frac{Y_t}{Q_t + Y_t}\right)P_0Y_0\int_t^T e^{((1-\pi)\mu-\rho)s}ds - (\beta - g(\beta))\left(1 + \phi\frac{V_t}{X_t + V_t}\right)w_0V_0\int_t^T e^{((1+\phi)\nu-\rho)s}ds} \tag{57}$$

Taking the derivative of equation (56) with respect to time and setting the result equal to zero yields (see Appendix):

$$(1 - \pi)\mu = (1 + \phi)\nu \tag{58}$$

The steady state direction of technical progress is therefore:

$$\frac{g(\beta)}{\beta} = \frac{\dot{Y}/Y - \dot{V}/V}{\dot{Y}/Y} = \frac{\mu - \nu}{\mu} = \frac{\pi + \phi}{1 + \phi} \tag{59}$$

In steady state growth equilibrium, the direction of investment $g(\beta)/\beta$ is the same for all firms in the industry, and depends on elasticity parameters only. Note that it does not depend on the number of firms in the industry. The higher π, the more rationalization. Next consider the budget for investment M of a firm in an industry on a steady state growth path. The time derivative of equation (57) can also be set to zero, resulting in (see Appendix):

$$\mu = \frac{\rho}{1 - \pi}; \quad \nu = \frac{\rho}{1 + \phi} \tag{60}$$

[14] Alternatively, what follows can be interpreted as the planning process of a firm that calculates its investment plan assuming constant elasticities *and* a constant market share on markets for in- and outputs. It is then analysed under which conditions the firm plans a constant growth rate of output, a constant growth rate of variable input demand and thus a constant rate of productivity increase. It should be noted that planned steady state growth only turns into real steady state growth if expectations are fulfilled, if plans and expectations of all firms in the market are consistent.

Let us consider what these outcomes mean. The *planned* growth rate of the firm's output supply and input demand depend on the outcome of the optimization problem of equations (47) to (51). These planned rates may be higher or lower than μ and ν. Here μ could be called the required or '*warranted*' rate that would ensure a steady state growth of output and ν the '*warranted*' rate of input demand growth. Note that there is presently nothing in the model that ensures that the actual rates move to the warranted rates or *vise versa*, e.g. through a movement of capital goods or other prices or of the interest rate.[15]

First consider the output side, making the assumption that on the input side the planned growth rate of input demand equals the warranted rate. Equation (60) shows that the warranted growth rate of output μ is *at least* as large as the interest (or discount) rate. Thus, if π is zero and firms plan an optimal growth rate exactly equal to the interest rate ρ, then firms will plan a steady state growth path. In this case the marginal return of investment equals the interest rate at precisely that growth rate of output that ensures that next period an equal investment is optimal. If the optimal growth rate is below the warranted rate μ, then optimal investment and output growth are planned to decrease over the planning period, and if it is above, then investment and the output growth rate are planned to increase (to see all this, refer to equations (52) and (53)). The warranted rate μ is thus an *unstable* equilibrium rate.

The warranted growth rate of output is determined by ρ and π. As the interest rate rises ρ, the warranted rate μ is higher. The firm's optimum growth path needs to be on a higher level, for it to be profitable to keep up the same rate of growth. As π is larger (demand being less elastic), the warranted rate rises too, because growth must not only compensate for the costs of capital at a certain level to stay attractive at a constant level, but also for the fall in the output price. The closer π gets to unity, the higher the planned equilibrium rate of the firm must be to equal the warranted rate, and thus the more likely it is that firms will plan a decrease in expansionary investment over time. If $\pi \geq 1$, if the elasticity of output demand is smaller than unity, there can be no steady state growth of output at all, implying that planned output growth will certainly be slowing down to zero in the long run.

Now consider the input side, making the assumption that the output side is on the equilibrium growth path. On the input side the warranted rate ν is *at most* equal to the interest rate. If ϕ is high (the supply elasticity of variable inputs is low), the warranted rate is small. The consequences of a planned growth rate exceeding the warranted rate ν (something which will happen the more easily as input supply is more inelastic) are twofold. A planned growth rate of input demand that is larger than the warranted rate will induce firms to redirect investment toward rationalization (see equation (52)), and at the same time to reduce the total amount of investment (see equation (53)), at least if $\beta > g(\beta)$. If the former outweighs the latter effect, and even more so if $\beta < g(\beta)$, the consequence for the planned growth rate of input demand is that, because more investment in rationalization will depress the planned growth rate, it will move toward the warranted rate ν. The warranted rate ν is thus a *stable* equilibrium.

[15] This is not Harrod-Domar, but there are similarities; however, we are only considering a partial equilibrium model, where the interest rate and the price of capital have an equilibrating function, but where interest and capital price movements are exogenous.

3.7 A model simulation of competition

To illustrate the model of section 6.2 a simulation program has been written. With the help of this program one can get an idea of the development of an economy where firms plan their actions in the way described. A simple simulation experiment will be presented here, to illustrate the general features of the model. The results of the simulation experiments are pictured in ten graphs below.

3.7.1 The setup of the simulation experiments

The complete model consists of equations (47) to (51). For the simulations below, the following groups of assumptions were made:

1. For the equations $g(\beta)$ and $h(M)$ simple specifications have been chosen, which fulfil requirements (2) and (1) respectively: $g(\beta) = a_0 + a_2\beta^2$, where $a_0 > 0, a_2 < 0$, and $h(M) = \delta_0 M^{\delta_1}$, where $\delta_0, \delta_1 > 0$ and $\delta_1 < 1$.

2. The firm is assumed to know the current value of the inverse price elasticities π_t and ϕ_t, which characterize the state of the markets for output and variable inputs at the moment of planning t. These elasticities change over time, as total supply of output or total factor demands change. This change of elasticities is derived from assuming a stable demand curve for output and a supply curve for variable inputs. The firms are assumed not to 'know' the global shape of these demand and supply curves, only to know the present values of the elasticities and use these as their expected values for the future. If the demand for output is related to its price as $\overline{Q}(P) = \mu_0 - \mu_1 P^{\mu_2}$, where $\mu_0, \mu_1, \mu_2 > 0$, and the supply of variable factors is related to its price by $\overline{X}(w) = \nu_0 + \nu_1 w^{\nu_2}$, where $\nu_1, \nu_2 > 0$, then $\pi = \mu_2^{-1}\left(\mu_0 / \left(\mu_1 P^{\mu_2}\right) - 1\right)$ and $\phi = \nu_2^{-1}\left(\nu_0 / \left(\nu_1 w^{\nu_2}\right) + 1\right)$.

3. Every firm makes its plans for future investments, given its expectations of actions of the other firms. What is relevant to the firm is the future aggregate production of all other firms and the future demand for inputs of all the others. For both the firm has to formulate expectations. It is assumed here that every firm simply extrapolates the current growth of both product supply and factor demand of the other firms into the future. The firm is not assumed to have all the necessary information freely at its disposal, nor to have the calculating capacity, to forecast simultaneously a mutually consistent set of the optimal strategies of all its competitors, conditional upon each other. Rather we assume that the firm applies a rule of thumb (a routine), making use of aggregate data on the present growth of markets. The current growth rate of competing supply $\dot{Q}_t / Q_t$ and demand $\dot{X}_t / X_t$ are calculated and used to predict future growth of competing supply and demand all the way to the planning horizon.

4. To concentrate on the dynamics of competition, the function $h(M)$ has been left independent of technological level τ and of spill-over effects from the technological frontier

θ, mentioned in section 5 above. The consequences of introducing these effects to the simulations are quite straightforward: a decrease in the speed of progress of the most advanced firms and a technological catching up of firms lagging behind in technology.

5. The following parameters were assumed: for $g(\beta)$ we took $a_0 = 1$, $a_2 = -1$; for $h(M)$ we took $\delta_0 = .005$, $\delta_1 = .5$. Furthermore, it was assumed that at the start of the process the situation on the output market was characterized by $\pi_0 = .2$ (i.e. the output elasticity is -5), and that the input market was characterized by $\phi_0 = .2$ (i.e. an input elasticity of 5). This signifies that the market for this output is still very 'young', and that the firms in the experiment are struggling to conquer this new market. Both the demand curve for output and the supply curve for variable inputs were assumed to be linear: $\mu_2 = 1$, $v_2 = 1$. Given prices for output and inputs at the start, plus aggregate output and aggregate factor demand at the start, the other parameters of the demand and supply curve can be calculated. They are in fact only scale parameters, since the important shape characteristics are determined by the four parameters above. The result was: $\mu_0 = 648$; $\mu_1 = 54$; $v_0 = -378$; $v_1 = 52.5$.

6. These last parameters were calculated under the assumption that at the start of the process the economy has the following features. At time 0 there are six firms, three large and three small. The large firms each produce 20 and the small each 16 units per period. The three firms of each scale operate with different variable factor productivities: 1, 8/7 and 4/3 respectively. Thus we start out from the following situation:

Table 4: Starting values, simulation experiment with six firms

firm	*output*	*variable inputs*	*productivity*
1.	20	20	1
2.	20	17.5	8/7
3.	20	15	4/3
4.	16	16	1
5.	16	14	8/7
6.	16	12	4/3

The price of output at time 0 is $P_0 = 10$ and the price of variable inputs $w_0 = 9$. The discount rate $\rho = .15$. Firms are assumed to plan twenty periods ahead and to replan every period. Thus in every period, a new plan for twenty periods is drawn up and immediately implemented. The simulation has run for 500 time periods. Over this time, the industry has moved along the demand and supply curves, ending up with a $\pi_{500} = 5.236$ and a $\phi_{500} = .058$.

The simulation program optimizes firm's plans, using equations (52) and (53). For starting values for time paths of β and M from the present period to the time horizon, using the above assumptions for estimating the moves of the competition, time paths for the development of production Y and variable input requirements V are calculated. These are used to calculate the expressions on the right hand side of (52) and (53). From these, values β and M for every period are calculated. In the optimum, these should be equal to the starting values of β and M. If they are not, the cycle is repeated with the new time paths for β and M as starting values, until convergence is reached.

The parameters and starting values had to be chosen such that the simulation program would easily converge to the optimal planning paths for the firms. The program tended not to converge when values for $h'(M)$ got too small. When $h'(M)$ gets very small, small changes in this derivative tend to go with large variations in investments M, causing instability in the optimization routine. To avoid this, the number of periods over which the simulation proceeds, 500 in total, has been made rather long, but $h(M)$ has got parameters such that progress goes slowly. The curvature of the investment function $h(M)$ is rather sharp with $\delta_1 = .5$, such that moderate progress is not expensive but large leaps are (an investment of 1 can produce at most .5% growth, an investment of 4 produces at most 1% growth, and so on). This makes larger firms divert not too fast from smaller firms. The time horizon was set at 20 periods. A longer planning period would induce firms to plan higher investments towards the end of the horizon, causing the instability in the optimization routine. However, it is not the time scale *per se* that is of interest, but the movements of firms relative to each other. These are adequately represented in the following graphs. Changes in parameters tend to change slopes of curves, but not their position relative to each other. After 500 periods, as shown in the graphs below, the industry is close to a stationary path. Extension of the time period only appeared to lead to a further decline in investment and stabilization of production and variable factor demand.

3.7.2 The results of the simulations

The results from the simulation experiment are depicted in Figures 1 to 10 below. Figure 1 illustrates the growth of production per enterprise. All firms grow over time, to between two and eight times their size at the start. The largest absolute growth is accomplished by the largest most efficient firm 3. Note that firm 3 grows fastest, but stops growing first. Since this firm has the largest market share, it internalizes a price decrease following upon output expansion most: its present revenues fall most as a consequence of a price drop. Therefore it is the first firm to stop causing a decrease in output price. Smaller firms continue to expand after firm 3 has stopped to grow. This causes firm 3 to loose market share, but because the inverse price elasticity of output π also increases further as other firms expand, firm 3 does not recommence expansion. As time continues, more firms stop to grow. The end of the process is a stable division of the market between the six firms, with shares of unequal size (see Figure 2). Over the course of the process, market shares have first diverged, but later converged again. The smaller firms have lost somewhat in the end, and the largest, most efficient firm 3 has gained some, but firms 1 and 2 have not moved substantially.

The demand for variable inputs, pictured in Figure 3, at first decreases for all firms, as they try to increase productivity. Soon the more efficient firms start to expand production rapidly. Firms 3 and 6, which grow fastest, absorb an increasing share of the supply of inputs, such that they, although being most efficient at the start, consume more inputs than their respective competitors which were of the same size at the beginning. The share in the total of firm 3 increases to over one and a half times its value at the start (see Figure 4). As time progresses and firms stop to expand, input demand decreases since firms redirect their investments again toward rationalization. Only the variable input requirements of the least efficient firm decreases from the beginning. Because smaller firms expand longer and large firms invest more effectively in raising variable factor productivity, smaller firms are least efficient in the end and larger firms most

Figure 1

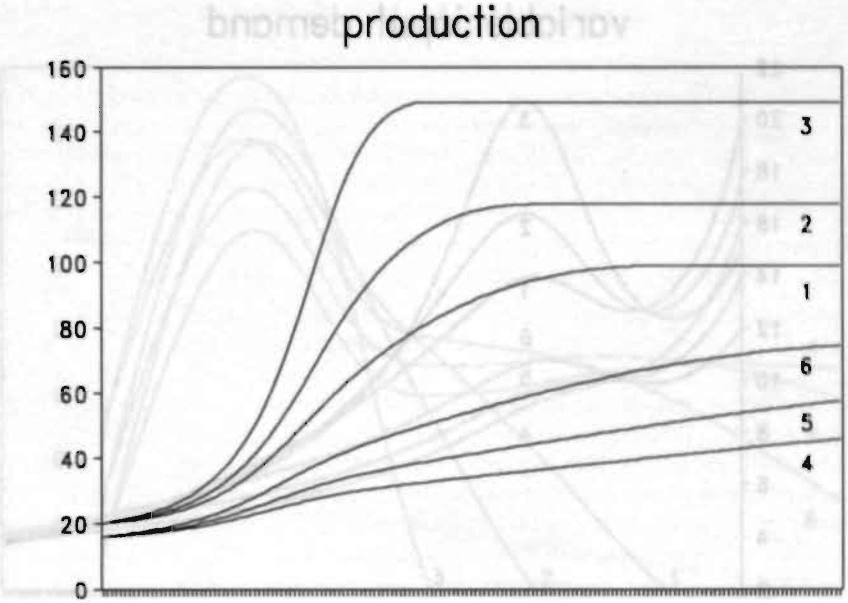


Figure 2

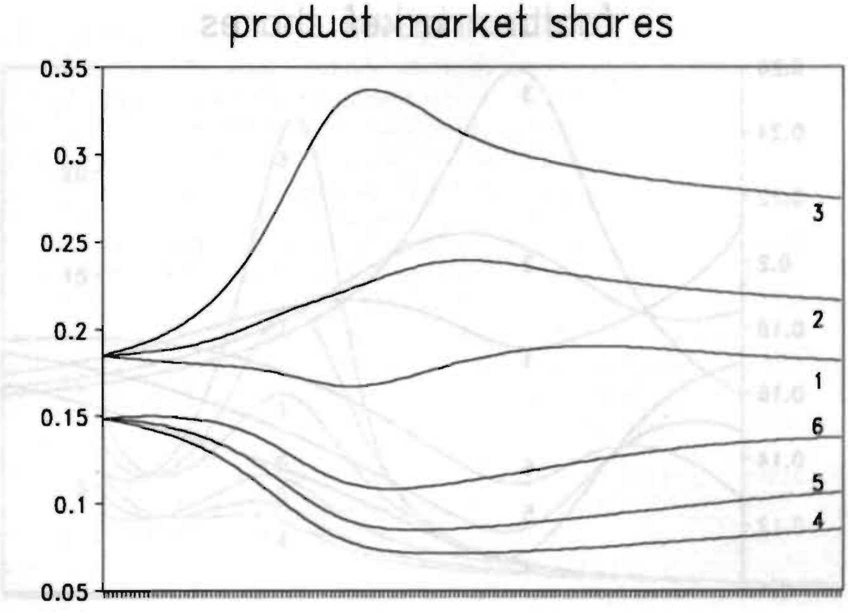


Figure 3

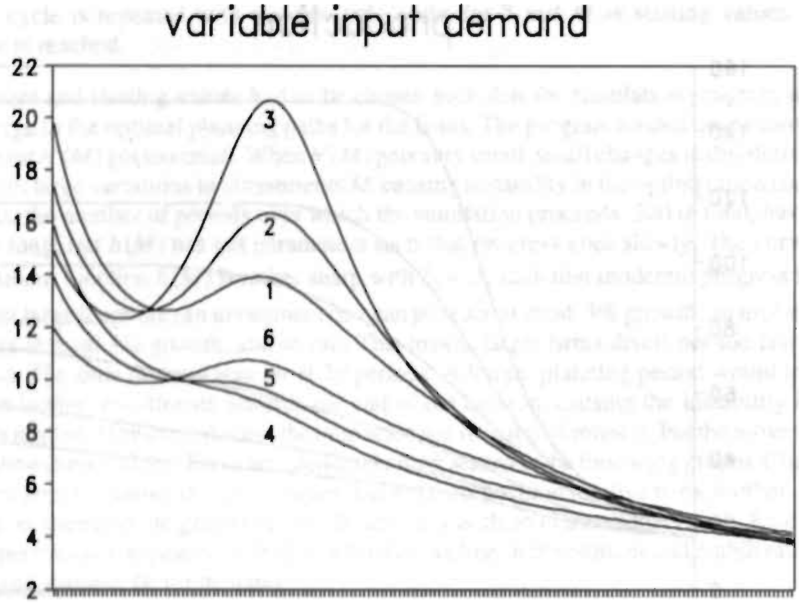


Figure 4

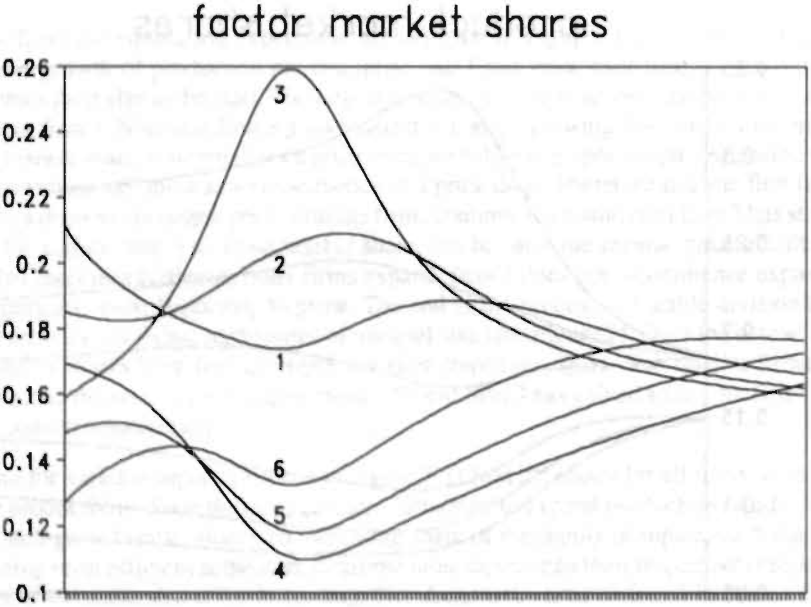


Figure 5

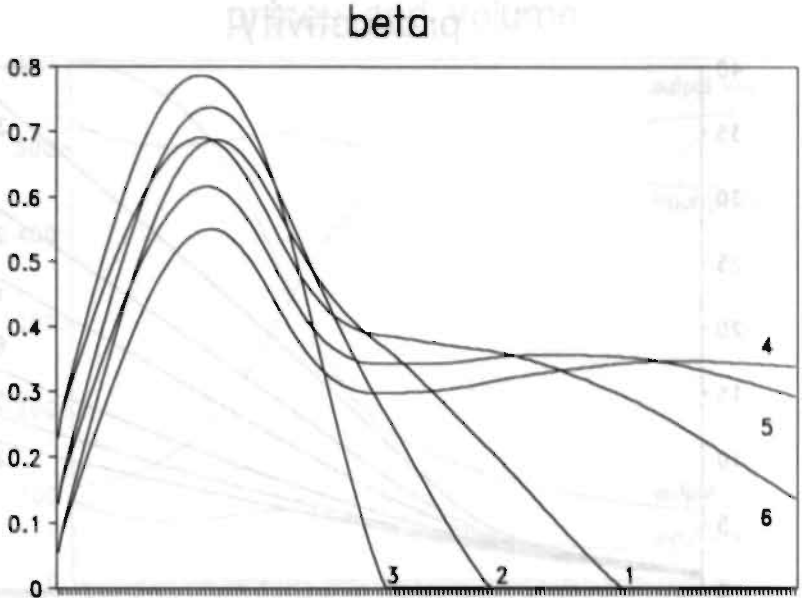


Figure 6

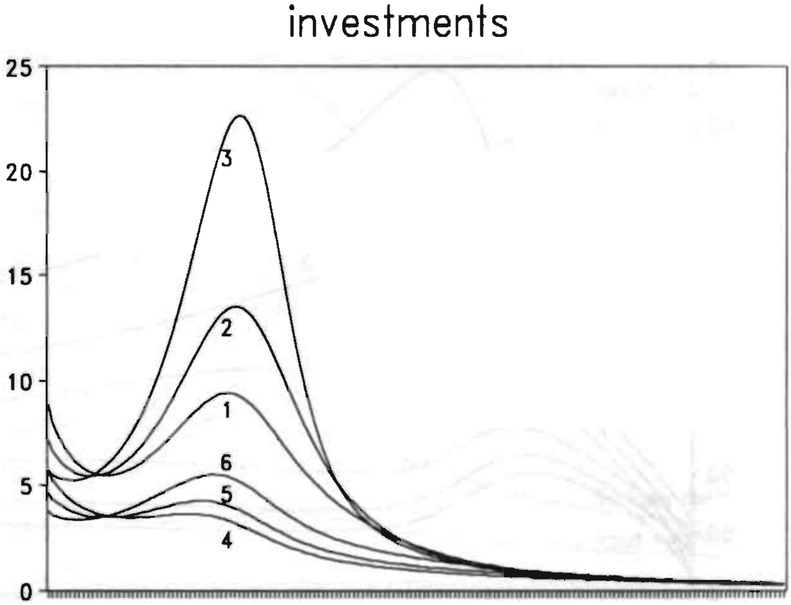


Figure 7

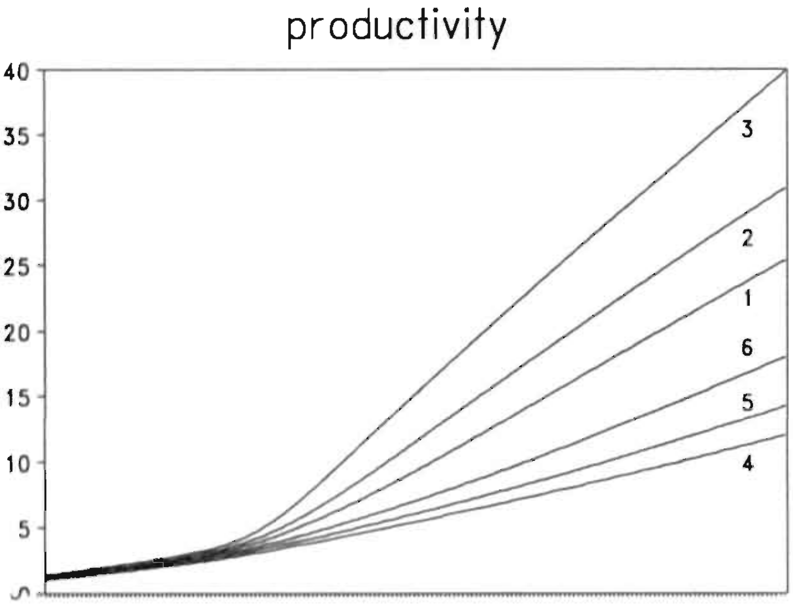


Figure 8

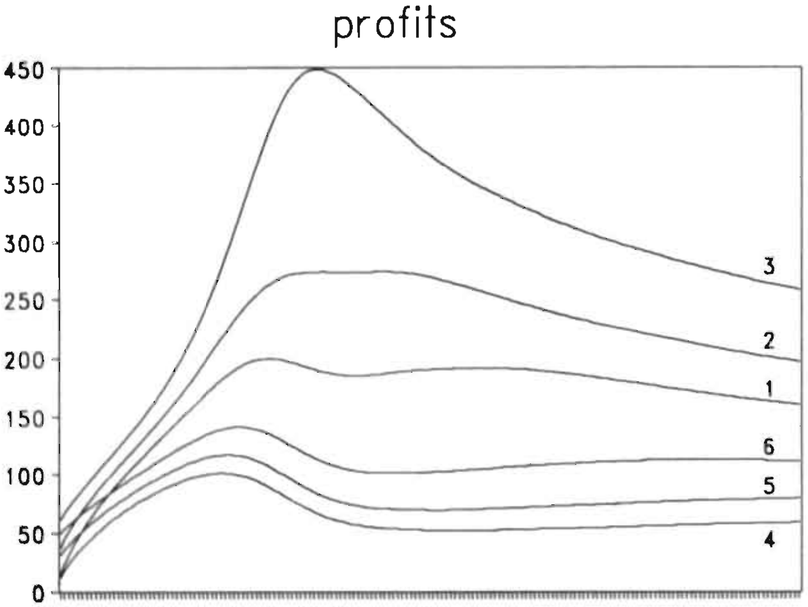


Figure 9

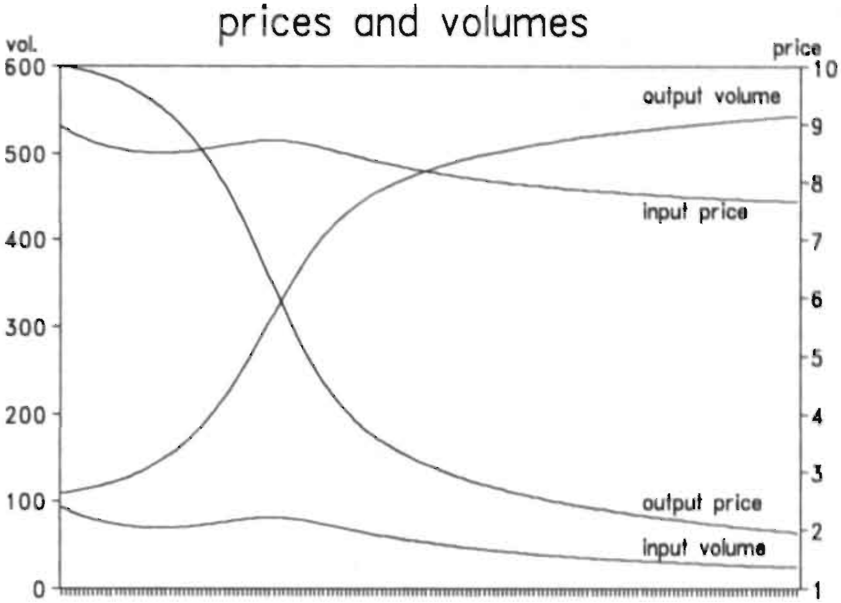
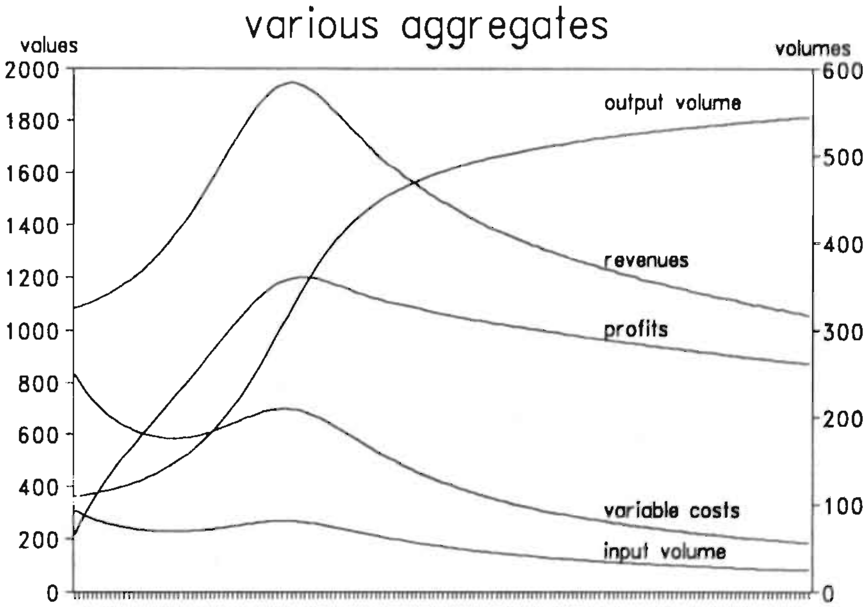


Figure 10



efficient, such that the smaller firms absorb the largest shares on the market for variable inputs, and the larger firms the smallest shares. From output and input data variable factor productivities can be calculated, as depicted in Figure 7. Note that firm 2 surpasses firm 6 and firm 1 surpasses firms 6 and 5 relatively quickly in productivity. Over 500 periods productivity grows to very high levels, because no limits have been put to the effectiveness of investments.

Figures 5 and 6 show the directions and sizes of investment by the six firms that have brought these developments about. Firm 1 is the biggest investor at the start, but invests predominantly in rationalization (a low β and a high $g(\beta)$). This firm is quickly overtaken by firms 3 and 2, which soon begin to invest in expansion. After initial rationalizations, all firms invest heavily in output expansion, as long as the market for output is not yet saturated: β's are growing. As markets become flooded, firms gradually lower β and increase $g(\beta)$, directing their investment streams, which are getting quickly smaller in volume now, toward rises in productivity. In the first periods, the same things happen among the smaller firms as among the larger firms, but on a smaller scale: firm 6 takes the same direction as 3, firm 5 follows 2 and firm 4 follows the direction of 1. As the larger firms stop expanding, or just before, there is an inducement to the smaller firms to redirect investment again toward expansion. This could be due to the expectation formation mechanism of the model. As firms 3, 2 and 1 stop expanding so suddenly, firms 4, 5, and 6 readjust their a expectations of the growth of competing supply. They realize that, as competing supply is not growing any more, their own supply can be profitably expanded further than they planned some periods before. Towards the end of the 500 periods investment virtually ceases. The advantages of expansion do not outweigh the disadvantage of a price drop any more for any of the six firms, and a further increase in variable factor productivity does no longer outweigh the costs.

The next graph, Figure 8, displays the realization of the objective of the firms: profits (these are free profits, calculated as cash flow minus investments in that period).[16] As the market is conquered in the first periods, profits sharply diverge, and the largest and most efficient firm takes most advantage of the opportunities. However, even *before* total output of the most efficient firm reaches its peak, profits already start to drop. As expansion of the smaller firms brings prices down further and profits of the largest firms erode, the gap between the firms becomes smaller. Profits of the smallest firms even increase a little. By the time the industry has matured, spectacular profit levels have been eliminated, and firms settle at profits which differ from each other about as much as at the start of the process.

Finally there are two graphs, Figures 9 and 10, showing a number of aggregate variables. Figure 9 pictures aggregate production and aggregate variable input demand, the output price and the input price. The output price drops sharply as aggregate output grows about fivefold, but because the rise of productivity compensates for expansion of output, the price of the variable inputs shows little variation. Note that the aggregate output curve is sigmoid. Thus the model gives a diffusion curve of the final good which has the familiar S-shape. Figure 10 presents total revenue, total expenditures on variable inputs, total profits and average variable factor productivity, and once more aggregate production and aggregate variable input demand. Note again that profits are highest before the market has reached its largest volume, and long before productivity is at

[16] Investments are fully paid out of current cash flow. Therefore the remaining profits is the sum which can be withdrawn without affecting the profits generated in the future. It is the amount of revenue that can more profitably be put on a bank account against an interest of ρ then be reinvested in the firm. Alternatively one could say that at the time of investment, the firm makes a reservation of an amount M, to pay every period costs of capital rM from interest revenues.

its top. Average productivity increase accelerates when profits start to diminish. The pattern of high profits preceding large volumes is reminiscent of the product life-cycle theory. According to this theory, the largest profits are made during the stage of sharpest market growth, and the largest volumes are sold as the product matures and low cost mass production prevails. The overall patterns of investment in the experiment of industry development also accords with the product life-cycle theory. The first part of the cycle is characterized by expansion of the industry, and the last by rationalization of production, by increases in productivity.

3.7.3 Evaluation

Some restrictive features of the simulation experiment and the findings need to be stressed. First of all, it is assumed that technological opportunities are never exhausted, and that firms can accomplish the same growth rates for the same volume of investment at any moment in time. In terms of the model, in the above experiment it is supposed that $h(M, \tau) = h(M)$. Secondly, it is assumed that there are no spill-overs of technological knowledge and skills from the most advanced firms to the others. Relaxing these restrictions will cause the most advanced firms to slow down their progress and the least advanced firms to catch up. This will change the time scale of the process, the distances between the firms, the speeds of change and adjustment. However, the general pattern of the competitive struggle, the positions and moves of the firms relative to the others, is not likely to be changed. A third limitation is the restriction on entry: entry barriers are assumed to be effective from the very start of the process onward. Profits remain in the end on positive levels, because there is no firm able to compete them away. Only firms that would be able to enter with a smaller market share than the smallest of the firms currently in the market would be willing to expand output any further. It is not specified, however, how and at what cost a firm could become established in the market, nor how incumbents would react to the threat of entry. Conditions for entry and costs of entry are outside the scope of the model. It is an empirical matter to what degree threat of entry is an important phenomenon on developing markets and to what degree an effective entry barrier limits the descriptive power of the model.[17] Fourthly, there is the assumption that productive capacity has an infinite technical lifetime. Relaxing this assumption would only introduce a downward trend in capacity, which has to be compensated for by extra investment (maintenance), but is also not likely to change the outcomes qualitatively.[18] Fifthly, it is assumed that both the price of investment goods and the interest rate are exogenously determined and constant, and that there are no liquidity constraints. Sixthly, the assumed expectations formation mechanism may be restrictive. Rational expectations of competitors actions, given the current value of price elasticities, instead of the above type of adaptive expectations, may alter the outcomes, but the consequences of introducing this assumption are hard to predict. Maybe smaller firms will invest and expand more aggressively at the start, maybe larger firms will pre-empt smaller firms. Multiple Nash equilibria cannot be excluded. Finally, firms' expectations about, or knowledge of, the global shape of the relevant demand and supply curves might affect the outcomes. The degree of realism of the assumptions on the way firms evaluate their competitors' likely behaviour and the characteristics

[17] Empirical studies suggest that there is a lot of entry in many markets, but little substantial post-entry market penetration [Geroski, Gilbert and Jacquemin (1990)].

[18] Economic lifetime and economic obsolescence, by the way, are not relevant concepts here: firms are assumed to invest in the most profitable way, and whether this is by replacing obsolete equipment, by adding new equipment or by rebuilding and improving old equipment, does not need to be specified, as long as the total effectiveness of investment M can be captured by the function $h(M)$.

of the demand side of the market can only be evaluated empirically. However, if information is incomplete and costly, and if evaluation is tedious, then the above assumptions might approach common practice rather closely.

Given the limitations of the scope of the experiment, the main conclusions may now be summarized. The simulations show how a number of firms of different size and efficiency struggle to conquer a new market. The following features stand out:

1. Over the process of competing over this market, output soars and the price of the product falls, but, due to technological progress, aggregate input demand and the input price change far less.
2. Market shares at first diverge, but later converge again, because the largest firms stop to grow while the smaller firms still keep on expanding. In this model, inertia on the part of large firms can be rational, when output is price-inelastic.
3. Like market shares, profits also first diverge and later converge again. In market shares and profits, there is falling behind and catching up. In productivity, however, lost grounds never tend to be made up for.
4. Small firms compete most effectively with large firms when markets start to be saturated, because large firms can only rely on rationalization to protect their profit levels, whereas small firms can still expand profitably.
5. Investment halts because profitable investment opportunities have been eroded, despite the fact that profits are positive. Even without retaliatory action of other participants in the market being expected, it is not profitable for any of the firms to expand any further. Thus the industry moves to a status quo, in which production continues unaltered ad infinitum, and in which firms keep on being different in market share, efficiency and profitability.
6. The time paths of the aggregate variables, describing the development of the industry are in accordance with elements from the product life-cycle theory. It has been found that the industry moves from an expansionary stage to a stage of rationalization, and that the largest profits are made in the periods of sharpest market growth.
7. The model generates a sigmoid diffusion curve. At first firms expand supply at an increasing rate, as they increasingly exploit scale economies; later, as markets saturate and price drops start to exert an increasingly negative influence on profits, firms increase supply at a decreasing rate.

3.8 Conclusions

In this chapter we developed a model of firm planning behaviour, assuming that a firm maximizes the present value of future profits, given a number of constraints. The most important constraint is the fact that the firm has to start out at the beginning of next period from where it will end up at the end of this period: the firm is tied to its own history. The firm is already established, with its managers and workers, its capital stock, bank credits, distribution channels, technological assets, etc., and any further move starts from this point. The firm is inflexible, in the sense that any change in its operations has a price. Change is costly to the firm, because it involves information gathering, research, discovery and learning, decision making, communication and persuasion, material investment, product development, technical adjustment and overcoming inertia, building up new routines and mastering new skills. All these processes are costly, in terms of money, but most importantly, in terms of time.

The firm makes a distinction between variable and fixed inputs. It can invest, add to the fixed inputs, to realize both capacity growth and growth of productivity of the variable inputs. Possibilities are limited by the nature of technological progress: there are decreasing returns to investment in any period. Depending on the stage of development of the technological trajectory that the firm exploits, there can be increasing or decreasing returns to investment over time. The firm's possibilities are also limited by supply of variable inputs and by demand for its output. The firm accounts for this in drawing up its plans, through its estimation of price elasticities on relevant markets, and by forecasting its competitors moves. Thus the firm plans its course of development with one eye on the market, taking account of input and output price elasticities and the likely development of its market share, and with the other eye on its production possibilities, its options for expansion and rationalization through the introduction of new technology.

This model of the firm leaves a number of relationships unspecified. To arrive at a description of a specific firm, more precise hypotheses have to be added to the present framework. First of all, one needs to specify which inputs are variable, and what their price is. Then the shape of $g(\beta)$, the innovation possibility frontier for given an amount of investment, and the nature of the investment function $h(M, \tau)$ must be specified. Technology spill-overs may be added. Finally, the firm's perceptions of its competitors, its expectations formation mechanisms, must be clarified. Firms form expectations on the inverse elasticities π and φ, and on the growth of supply of the final good and the demand for variable factors by competitors, $\dot{Q}$ and $\dot{X}$ respectively. The latter include expectations on entry and exit. These issues have been left open here and the analysis has concentrated on the general features of the model. The conclusions reached therefore hold for a fairly wide range of specific models.

First of all, this model was used to analyse the process of investment planning of the firm. One finding is that, if input and output markets are perfectly competitive, firms tend to plan an amount of investment which is permanently growing over time and tend to direct their investment ever less toward factor productivity improvements and ever more toward capacity expansion. Returns to investment in productivity growth erode, because variable factors of production constitute a decreasing share of total costs of production.[19] By contrast, returns to investment in expansion increase, because of the cumulative effect of technical change: the same investment produces an ever larger expansion of output. However, when markets are not perfectly competitive (expressed by falling price elasticities as markets mature), this conclusion no longer holds: firms will decrease investments in the long term, if the demand elasticity of output rises. On the one hand, returns to investment in productivity erode, for the same reason as before. On the other hand, returns to investment in expansion also erode, because expansion of output forces prices down. Expansion of output proceeds until extra revenue from more sales cannot compensate any more for the loss of revenue from the current level of sales, due to the drop in output prices. There is an intermediate case, where the price elasticity of output demand takes on a constant value, larger than unity. If that is the case, there is a direction of investment $g(\beta)/\beta$ and an amount of investment M, that could sustain a steady state growth of output and input demand. The steady state growth rates of firm's output supply and input demand are functions of the interest rate and the elasticity of output demand and input supply respectively. There is no mechanism specified, however, that leads firms toward the steady state path and in addition this path, once attained, turns out to be unstable.

[19] This is not related to the Marxian idea of the tendency of the rate of profit to fall. There is no equivalent in the model above to Marx' assumption of a constant degree of exploitation of variable factors of production, which is at the basis of his theorem.

Another finding concerns the firm's investment planning and technological expectations. By definition technological change occurs if an investment has a cumulative effect, in the sense that the consequence of an investment now is that an investment of the same size in the next period has a larger effect. Firm planning was considered under four different assumptions: a) technical opportunities are such that investment has a cumulative effect *and* the firm also expects its investments to lead to technical progress; b) there are no opportunities for technical progress, *but* the firm expects its investments to lead to technical change; c) there are opportunities, *but* the firm does not expect them; d) there are no opportunities *and* they are not expected either. It was shown that technological change, expected or not, always leads to higher output, higher variable factor requirements and higher variable factor productivity. Also it turned out that expecting technological change leads to higher output and variable factor requirements than expecting no technological change, whether it actually occurs or not. Variable factor productivity, however, may be *lower* when technological change is expected than when it is not. This occurs when firms expecting technical progress expand too much, as they count on learning effects, on reaping dynamic economies of scale in future technical progress.

Beside expectations about technological opportunities, firms form expectations of competitor behaviour. The reaction of a firm expecting a higher growth in competing output supply depends on the price elasticity of demand for output and the firm's market share. Firms will plan retaliatory expansion if they expect output demand to be inelastic and to have a large share of the market in the future; they will plan to limit their expansion if demand for output is elastic and the expected market share is small. A critical size above which firms retaliate was derived as a function of the elasticity of demand and the size of competing supply.

Secondly, the model was used to analyse the process of competition within an industry, both analytically and by means of simulation techniques. It may be noted that the model provides a framework which can accommodate for a wide range of real world phenomena, that demand more ad hoc assumptions in other frameworks. An industry with firms that act along the lines of the model does not necessarily end up in monopoly, if increasing returns over time are prominent. Different market structures, industries made up out of small and large firms, are not only possible as transitory phenomena, but can be stable over prolonged periods. New technology does not diffuse immediately, small firms do not necessarily loose in the competitive struggle, nor do technologically backward firms necessarily go bankrupt. There is a long term tendency to direct investments in technical change toward productivity growth, but in the short term firms may move away from this direction. Investments in the exploitation of a specific technological paradigm will cease in the end, but in the short term they can decrease as well as increase. The model of the firm is a general building block to build various types of economic structures with it. This is the case, even though firms, given their access to information, plan rationally, with an infinite time horizon, though they operate in a deterministic world, and though technological opportunities, as expressed by $g(\beta)$ and $h(M,\tau)$, are the same for all firms.

A result from the analysis are the qualifications that can be put forward, concerning the trade off between static and dynamic efficiency. Standard analysis holds, that within a static context, perfect competition leads to optimal allocation of resources. However, it is also often stated [see e.g. Scherer (1980), Kamien and Schwartz (1982)], that perfect or nearly perfect competition is detrimental to innovative effort and technical progress, and thus leads to dynamic suboptimality, if intertemporal efficiency is considered. A debate centres around the relative merits of monopoly versus some monopolistic or oligopolistic form of competition for aggregate growth. Within the confines of the model analysed above, it has been shown that, even in a dynamic framework, *more* competition can be superior to *less* competition. If a market of given size is assumed to be divided among a number of identical firms, the sign of the relationship between market concentration and the amount of investment of each firm separately depends on the elasticity of output demand. If demand is elastic, the more firms are in the market, the lower the investment budget of each competitor is. However, if output demand is inelastic, then an increase in the number of firms may induce each of them to invest more. A sufficient condition for investment to grow with the number of competitors was derived above.

This result can be understood as follows. Suppose that the industry currently produces a given volume of output and that the demand curve for output is fixed and relatively inelastic. If the total costs of expansion of the market would be the same for either one monopolist or for n competing firms, the monopolist would tend to expand less than the n competitors. This occurs, because the monopolist takes the full repercussions of the decrease in price on his decisions into account; the effect of the price decrease is fully internalized. For competitors that have a smaller share of the market, however, the price decrease is only partly caused by the firm's own expansion of output, and for the rest it is an externality, produced collectively by all other firms in the market. Due to this externality, the resulting price is not optimal for the collective of competitors; if the competitors would collude and coordinate their actions, they would restrict output expansion to support the price. However, because they compete, they are caught in a type of prisoner's dilemma, since without coordination, it is rational for each of them to expand output further than would happen with coordination. The consumer side benefits from competition, not only in a static but also in a dynamic world.

Another result of the analysis of competition was the light shed upon the relationship between technological change, market structure and firm size in a dynamic context (the Schumpeterian hypotheses). Firm size (absolute size) and market share (relative size) were shown to influence investment in growth simultaneously but in opposing directions: firm size positively and market share negatively. Although these two variables are usually highly correlated, their influence on investment behaviour should not be considered in isolation. The importance of firm size and market structure as determinants of the speed of progress varies over the life-cycle of the market or industry. As the output price falls and the elasticity of demand rises, the dominance of the firm size factor decreases and the influence of the market share argument increases: in a developing market firm size dominates market structure, in a mature market *vise versa*.

A next result worth mentioning, is that the model sheds some light on investment activities in relationship to the life cycle development of a product, or on a higher level of aggregation, to the maturity of the industry. As products are longer on the market and go through different stages of the life cycle, as gradually industries mature, markets get increasingly saturated. This causes the price elasticity of output to decline: a percentage increase of output supply causes a larger percentage fall in the price. Similarly, if firms buy their inputs in a mature market, relative changes in input requirements elicit larger relative changes in factor prices. Firms recognizing this will calculate their investment plans in an emerging market or industry, assuming elastic in- and output markets (a low π and ϕ) and in a mature market or industry, assuming inelastic markets (a high π and ϕ). Therefore, firms will invest a larger share of their cash flow in an emerging industry and direct investment relatively more toward expansion. Conversely, firms will invest modestly in a mature industry and strive predominantly for rationalization, for bringing down the costs of production. Thus the beginning of the life cycle is characterized by expansion, possibly supported by product differentiation, whereas competition in the later stages of the product life cycle is characterized by price competition. Furthermore, simulations showed that not only investment in the model develops according to a frequently observed and characteristic pattern over the life cycle, but also industry profits: aggregate profits increase as the market develops, but start to decline before output volumes reach their maximum. Profits rise in the model as long as output expansion compensates for price falls, and then start to fall.

A final result to mention is that the model generates a sigmoid diffusion curve of output. As a market opens up, first firms expand supply at an increasing rate, as they increasingly exploit scale economies. As the market saturates and the output price falls, thereby putting profits increasingly under pressure, firms increase supply at a decreasing rate. In summary, it may be noted that there are a number of aggregate phenomena produced by the model in the simulation experiment, that resemble regular patterns well-known from empirical research. The fact that the simulations, which represent the interactions of units modelled at the micro level only, generate familiar aggregate phenomena may be taken as support for the assumptions that are at the basis of the micromodel.

This completes the theoretical exposition of the model of firm planning. The model comes close to the criteria formulated for a model of economic development in section 2.2 of chapter 2: the firm invests in the production of technology, in the determination of its own production constraints. In the framework represented in Table 1 of chapter 1, this model may be situated somewhere in the middle between mainstream and evolutionary thought. In terms of this table, the following characteristics are important. First of all, technological change is a main determinant of the development of the firm and the industry. Technology changes endogenously, but only incrementally. Secondly, although the firm optimizes an objective function, its objective is closely tied to its present situation: investment is the price of a change of existing routines. Thirdly, there is no uncertainty, but the fact that the firm is tied to its past technology and scale expresses that change is complex and costly. One reason why change is costly is because to find and use information on how to change is costly. Fourthly, given available information, the firm takes optimal decisions, but these are not necessarily consistent nor inconsistent with the decisions of other firms in the market. Finally the development of the firm is evidently path dependent.

Mathematical appendix

Derivations section 3

Analysis of the basic model:

We calculate the Hamiltonian and take derivatives to the instruments β and M, and to the variables Y_t and V_t.

$$H = e^{-\rho t}(PY_t - wV_t - M) + \lambda_1 Y_t \beta h(M) + \lambda_2 V_t(\beta - g(\beta))h(M) \tag{A1}$$

Differentiating yields:

$$\frac{\delta H}{\delta \beta} = \lambda_1 Y_t h(M) + \lambda_2 V_t(1 - g'(\beta))h(M) = 0 \quad \Leftrightarrow \quad g'(\beta) = 1 + \frac{\lambda_1 Y_t}{\lambda_2 V_t} \tag{A2}$$

$$\frac{\delta H}{\delta M} = -e^{-\rho t} + (\lambda_1 Y_t \beta + \lambda_2 V_t(\beta - g(\beta)))h'(M) = 0 \quad \Leftrightarrow \quad h'(M) = \frac{e^{-\rho t}}{\lambda_1 Y_t \beta + \lambda_2 V_t(\beta - g(\beta))} \tag{A3}$$

The change in the costate variables can be obtained by differentiating the Hamiltonian with respect to Y_t and V_t.

$$\frac{\delta H}{\delta Y_t} = e^{-\rho t}P + \lambda_1 \beta h(M) = -\dot{\lambda}_1 \tag{A4}$$

$$\frac{\delta H}{\delta V_t} = -e^{-\rho t}w + \lambda_2(\beta - g(\beta))h(M) = -\dot{\lambda}_2 \tag{A5}$$

This can be written as:

$$Y_t \dot{\lambda}_1 + \dot{Y}_t \lambda_1 = -e^{-\rho t} P Y_t \tag{A6}$$

$$V_t \dot{\lambda}_2 + \dot{V}_t \lambda_2 = e^{-\rho t} w V_t \tag{A7}$$

These equations can be integrated, taking account of the condition that $\lambda_1(T) = \lambda_2(T) = 0$, which says that the shadow prices of changes in output or variable input demand at the planning horizon (or beyond) equal zero:

$$Y_t \lambda_1 = \int_t^T e^{-\rho s} P Y_s ds \tag{A8}$$

$$V_t\lambda_2 = -\int_t^T e^{-\rho s} w V_s ds \tag{A9}$$

We can substitute equations (A8) and (A9) in (A2) and (A3) respectively. Solution values of β and *M* must satisfy at all times t:

$$g'(\beta) = 1 - \frac{\int_t^T e^{-\rho s} P Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} \tag{A10}$$

$$h'(M) = \frac{e^{-\rho t}}{\beta \int_t^T e^{-\rho s} P Y_s ds - (\beta - g(\beta)) \int_t^T e^{-\rho s} w V_s ds} \tag{A11}$$

Stability of the solution $g(\beta) = 0$:

$$\frac{dg'(\beta)}{dt} = g''(\beta)\frac{d\beta}{dt} \tag{A12}$$

Since by construction $g''(\beta) < 0$, we have from equation (9) a value of $\beta > 0$ if:

$$PY_t + (g'(\beta) - 1)wV_t < 0 \quad \Leftrightarrow \quad g'(\beta) < 1 - \frac{PY_t}{wV_t} \tag{A13}$$

That this is always the case follows from (A10) and:

$$\frac{\int_t^T e^{-\rho s} P Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} > \frac{\int_t^T e^{-\rho s} P Y_t ds}{\int_t^T e^{-\rho s} w V_t ds} = \frac{PY_t}{wV_t} \tag{A14}$$

The inequality holds because the first expression is the ratio of total revenue and total variable costs when the firm follows an optimizing strategy and the second expression is the same ratio in case the firm takes no action. The optimizing strategy is always at least as profitable as doing nothing. Thus the firm will always end up spending all its investments on capacity expansion.

Derivations section 5

Analysis of the model with depletion of technological opportunities:

From the Hamiltonian, five first order conditions can be derived:

$$H = e^{-\rho t}(PY_t - wV_t - M) + \lambda_1 Y_t \beta h(M,\tau) + \lambda_2 V_t(\beta - g(\beta))h(M,\tau) + \lambda_3 h(M,\tau) \tag{A15}$$

Differentiating yields as before:

$$\frac{\delta H}{\delta \beta}=0 \quad \Leftrightarrow \quad g'(\beta)=1+\frac{\lambda_1 Y_t}{\lambda_2 V_t} \tag{A16}$$

$$\frac{\delta H}{\delta M}=0 \quad \Leftrightarrow \quad h_M=\frac{e^{-\rho t}}{\lambda_1 Y_t \beta+\lambda_2 V_t(\beta-g(\beta))+\lambda_3} \tag{A17}$$

$$\frac{\delta H}{\delta Y_t}=e^{-\rho t}P+\lambda_1\beta h(M,\tau)=-\dot{\lambda}_1 \quad \Leftrightarrow \quad \dot{\lambda}_1 Y_t+\lambda_1\dot{Y}_t=-e^{-\rho t}PY_t \tag{A18}$$

$$\frac{\delta H}{\delta V_t}=-e^{-\rho t}w+\lambda_2(\beta-g(\beta))h(M,\tau)=-\dot{\lambda}_2 \quad \Leftrightarrow \quad \dot{\lambda}_2 V_t+\lambda_2\dot{V}_t=e^{-\rho t}wV_t \tag{A19}$$

$$\frac{\delta H}{\delta \tau_t}=h_\tau(\lambda_1\beta Y_t+\lambda_2(\beta-g(\beta))V_t+\lambda_3)=-\dot{\lambda}_3 \tag{A20}$$

From (A17) and (A20) we can deduce:

$$h_M=\frac{e^{-\rho t}}{-\dot{\lambda}_3/h_\tau} \quad \Leftrightarrow \quad \dot{\lambda}_3=-e^{-\rho t}\frac{h_\tau}{h_M} \tag{A21}$$

This can be integrated to give:

$$\lambda_3=\int_t^T e^{-\rho s}\frac{h_\tau}{h_M}ds \tag{A22}$$

As before we can integrate (A18) and (A19) and substitute the result into (A16) to arrive at an expression from which we can evaluate the direction of technical progress.

$$g'(\beta)=1-\frac{\int_t^T e^{-\rho s}PY_s ds}{\int_t^T e^{-\rho s}wV_s ds} \tag{A23}$$

This expression is the same as in the original model, indicating that under this formulation the planned direction of technical change is not influenced by the history of investment, by the expenses on technical change in the past. Furthermore we can substitute (A22) and the integrals of (A18) and (A19) into (A17) to arrive at:

$$h_M=\frac{e^{-\rho t}}{\int_t^T e^{-\rho s}\left[\beta PY_s-(\beta-g(\beta))wV_s+\frac{h_\tau}{h_M}\right]ds} \tag{A24}$$

Derivations section 6

Analysis of the model with a varying price for output:

The Hamiltonian of this problem is:

$$H = e^{-\rho t}(P_tY_t - wV_t - M) + \lambda_1 Y_t\beta h(M) + \lambda_2 V_t(\beta - g(\beta))h(M) - \lambda_3 P_t\pi\frac{\dot{Q}_t + Y_t\beta h(M)}{Q_t + Y_t} \tag{A25}$$

Differentiating yields:

$$\frac{\delta H}{\delta\beta} = 0 \iff g'(\beta) = 1 + \frac{\lambda_1 Y_t}{\lambda_2 V_t} - \pi\frac{\lambda_3 P_t}{\lambda_2 V_t}\left(\frac{Y_t}{Q_t + Y_t}\right) \tag{A26}$$

$$\frac{\delta H}{\delta M} = 0 \iff h'(M) = \frac{e^{-\rho t}}{\lambda_1 Y_t\beta + \lambda_2 V_t(\beta - g(\beta)) - \lambda_3 P_t\pi\frac{Y_t\beta}{Q_t + Y_t}} \tag{A27}$$

$$\frac{\delta H}{\delta Y_t} = e^{-\rho t}P_t + \lambda_1\beta h(M) - \lambda_3 P_t\pi\frac{\beta h(M)Q_t - \dot{Q}_t}{(Q_t + Y_t)^2} = -\dot{\lambda}_1 \tag{A28}$$

$$\frac{\delta H}{\delta V_t} = -e^{-\rho t}w + \lambda_2(\beta - g(\beta))h(M) = -\dot{\lambda}_2 \tag{A29}$$

$$\frac{\delta H}{\delta P_t} = e^{-\rho t}Y_t - \lambda_3\pi\frac{\dot{Q}_t + \beta h(M)Y_t}{Q_t + Y_t} = -\dot{\lambda}_3 \tag{A30}$$

From the last two conditions we can derive:

$$V_t\lambda_2 = -\int_t^T e^{-\rho s}wV_s ds \tag{A31}$$

$$P_t\lambda_3 = \int_t^T e^{-\rho s}P_sY_s ds \tag{A32}$$

The last expression can be used to handle the derivative of the Hamiltonian to Y_t:

$$\dot{\lambda}_1 Y_t + \lambda_1\dot{Y}_t = -e^{-\rho t}P_tY_t + \lambda_3 P_t\pi\frac{\dot{Y}_tQ_t - \dot{Q}_tY_t}{(Q_t + Y_t)^2} = -e^{-\rho t}P_tY_t + \pi\frac{d\left(\frac{Y_t}{Q_t + Y_t}\right)}{dt}\int_t^T e^{-\rho s}P_sY_s ds \tag{A33}$$

This expression can be integrated:

$$\lambda_1 Y_t = \int_t^T e^{-\rho s} P_s Y_s ds - \pi \int_t^T \frac{d\left(\frac{Y_s}{Q_s + Y_s}\right)}{ds} \int_s^T e^{-\rho u} P_u Y_u du\, ds \tag{A34}$$

Manipulation of the last equation yields:

$$\lambda_1 Y_t = \int_t^T e^{-\rho s} P_s Y_s ds - \pi \left\{ \frac{Y_s}{Q_s + Y_s} \int_s^T e^{-\rho u} P_u Y_u du \right\} \Big|_t^T - \pi \int_t^T \frac{Y_s}{Q_s + Y_s} e^{-\rho s} P_s Y_s ds =$$
$$\left(1 + \pi \frac{Y_t}{Q_t + Y_t}\right) \int_t^T e^{-\rho s} P_s Y_s ds - \pi \int_t^T \frac{Y_s}{Q_s + Y_s} e^{-\rho s} P_s Y_s ds \tag{A35}$$

The expressions for $\lambda_1 Y_t$, $\lambda_2 V_t$ and $\lambda_3 P_t$ can be substituted in the formulas for $g'(\beta)$ and $h'(M)$. Some terms appear to cancel out and we get:

$$g'(\beta) = 1 - \frac{\int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds}{\int_t^T e^{-\rho s} w V_s ds} \tag{A36}$$

$$h'(M) = \frac{e^{-\rho t}}{\beta \int_t^T \left(1 - \pi \frac{Y_s}{Q_s + Y_s}\right) e^{-\rho s} P_s Y_s ds - (\beta - g(\beta)) \int_t^T e^{-\rho s} w V_s ds} \tag{A37}$$

Derivation of the more general result of section 2 goes along the same lines.

Analysis of the relationship between market concentration and growth:

Write equations (54) and (55) as follows:

$$g'(\beta) = 1 - \frac{c_1 \left(1 - \frac{\pi}{n}\right)}{c_2 \left(1 + \frac{\phi}{n}\right)} \qquad \text{where} \quad c_1 \equiv \int_t^T e^{-\rho s} P_s \overline{Q}_s ds \quad \text{and} \quad c_2 \equiv \int_t^T e^{-\rho s} w_s \overline{X}_s ds \tag{A38}$$

$$h'(M) = \frac{n e^{-\rho t}}{\beta \left(1 - \frac{\pi}{n}\right) c_1 - (\beta - g(\beta)) \left(1 + \frac{\phi}{n}\right) c_2} \tag{A39}$$

To trace the effect of a change in the number of firms n on the direction of investment $g(\beta)/\beta$, consider equation (A38). Remember that $\frac{dg'(\beta)}{dn} < 0$ implies that $g(\beta)/\beta$ changes negatively with n. The derivative with respect to n is:[20]

[20] Strictly the following expressions are an approximation, leaving out second round effects. Second round effects appear because as investment reacts to changes in n, there are changes in future growth of output and variable factor demand, which may induce further changes in investment budgets. In particular it is assumed here that the derivatives of Y_s and V_s, where $t \le s \le T$, with respect to the number of firms n is negligible. This would hold e.g. if the planning horizon is relatively short. If this is not the case, then 'perverse' effects or cyclical movements cannot be excluded.

$$\frac{dg'(\beta)}{dn} = -\frac{c_1}{c_2}\left\{\pi\left(1+\frac{\phi}{n}\right)+\phi\left(1-\frac{\pi}{n}\right)\right\}(n+\phi)^{-2} < 0 \quad \Leftrightarrow \quad n\left(\frac{1}{\pi}-\frac{1}{\phi}\right) < 2 \tag{A40}$$

This condition is more likely to be fulfilled as elasticity parameter π is large and φ is small. If the condition holds, if output demand is price inelastic and input supply is price elastic, then an increase in the number of firms will make firms redirect investment toward expansion.

To derive the effect of a change in the number of firms n on the size of the investment budget M, consider equation (A39) and note that $\frac{dM}{dn} > 0 \Leftrightarrow \frac{dh'(M)}{dn} < 0$. The derivative of $h'(M)$ with respect to the number of firms n is:

$$\frac{dh'(M)}{dn} = \frac{e^{-\rho s}\left(\beta\left(1-2\frac{\pi}{n}\right)c_1-(\beta-g(\beta))\left(1+2\frac{\phi}{n}\right)c_2\right)}{\left(\beta\left(1-\frac{\pi}{n}\right)c_1-(\beta-g(\beta))\left(1+\frac{\phi}{n}\right)c_2\right)^2} \tag{A41}$$

Therefore:

$$\frac{dh'(M)}{dn} < 0 \quad \Leftrightarrow \quad \frac{n-2\pi}{n+2\phi} < \left(1-\frac{g(\beta)}{\beta}\right)\frac{c_2}{c_1} \tag{A42}$$

Here $c_2/c_1 < 1$ if cash flow is to be positive. The condition is more likely to be fulfilled, if π is large, if the optimal direction of investment is toward expansion (β large) and n is small. Then a growth in the number of firms will induce an increase in investment M.

Analysis of the reaction of the firm to variation in expected competitor behaviour

Insert $Q_t = Q_0 e^{\psi t}$ in equation (53), take account of the fact that the firm assumes that $P_t = (Q_t + Y_t)^{-\pi}$, and consider the effect of a marginal change in the growth rate ψ of the expected volume of competing output Q_s on the volume of investment M. It follows that:

$$\frac{dM}{d\psi} > 0 \quad \Leftrightarrow \quad \frac{dh'(M)}{d\psi} < 0 \quad \Leftrightarrow \quad \frac{d}{d\psi}\left\{\int_t^T e^{-\rho s}\left(1-\pi\frac{Y_s}{Q_0e^{\psi s}+Y_s}\right)(Q_0e^{\psi s}+Y_s)^{-\pi}Y_s ds\right\} > 0 \tag{A43}$$

After taking derivatives, the last condition can be rewritten as:

$$\int_t^T e^{-\rho s}P_sQ_s s\pi\frac{Y_s}{Q_s+Y_s}\left\{(1+\pi)\frac{Y_s}{Q_s+Y_s}-1\right\}ds > 0 \tag{A44}$$

A sufficient condition for this inequality to hold is:

$$\forall t \le s \le T: \quad (1+\pi)\frac{Y_s}{Q_s+Y_s} > 1 \quad \Leftrightarrow \quad \pi > \frac{Q_s}{Y_s} \tag{A45}$$

It can be shown by applying a similar procedure to equation (52) that a sufficient condition for the redirecting of investment toward expansion, in reaction to an increase in the expectation of competing supply, is the same. Thus there is a threshold size of the firm, $Y = Q/\pi$, below which it backs out upon the threat of increased competition, decreasing investment and moving towards rationalization, and above which it retaliates by stepping up investment, especially in expansion.

Similarly consider the effect of a change in the expected growth rate of variable input demand on the volume and direction of investment:

$$\frac{dM}{d\xi} < 0, \quad \frac{d\beta}{d\xi} < 0 \quad \Leftrightarrow \quad \frac{dh'(M)}{d\xi} > 0, \quad \frac{dg'(\beta)}{d\xi} > 0 \quad \Leftrightarrow \tag{A46}$$

$$\int_t^T e^{-\rho s} w_s X_s s \phi \frac{V_s}{X_s + V_s} \left\{ (1-\phi) \frac{V_s}{X_s + V_s} - 1 \right\} ds < 0$$

Clearly these inequalities always hold, because $\forall t \leq s \leq T$: $-\phi < X_s / V_s$.

Analysis of the possibility of a steady state growth path:

Setting the derivative of equation (56) with respect to time equal to zero and simplifying, we get:

$$\frac{dg'(\beta)}{dt} = 0 \quad \Leftrightarrow \quad \int_t^T e^{\zeta_1 (s-t)} ds = \int_t^T e^{\zeta_2 (s-t)} ds \tag{A47}$$

$$\text{where} \quad \zeta_1 \equiv (1-\pi)\mu - \rho \quad \text{and} \quad \zeta_2 \equiv (1+\phi)\nu - \rho$$

Solving yields:

$$\frac{1}{\zeta_1 (s-t)} \left(e^{\zeta_1 (T-t)} - 1 \right) = \frac{1}{\zeta_2 (T-t)} \left(e^{\zeta_2 (T-t)} - 1 \right) \tag{A48}$$

This condition holds for all $0 \leq t \leq T$ if $\zeta_1 = \zeta_2 \quad \Leftrightarrow \quad (1-\pi)\mu = (1+\phi)\nu$.

Setting the time derivative of equation (57) equal to zero, we find:

$$\rho \left(c_1 \int_t^T e^{\zeta_1 s} ds - c_2 \int_t^T e^{\zeta_2 s} ds \right) = c_1 e^{\zeta_1 t} - c_2 e^{\zeta_2 t} \tag{A49}$$

$$\text{where} \quad c_1 \equiv \beta \left(1 - \pi \frac{Y_t}{Q_t + Y_t} \right) P_0 Y_0 \quad \text{and} \quad c_2 \equiv (\beta - g(\beta)) \left(1 + \phi \frac{V_t}{X_t + V_t} \right) w_0 V_0$$

Simplifying gives:

$$c_1 \left(e^{\zeta_1 t} - \frac{\rho}{\zeta_1 + \rho} e^{\zeta_1 T} \right) \left(1 + \frac{\rho}{\zeta_1} \right) = c_2 \left(e^{\zeta_2 t} - \frac{\rho}{\zeta_2 + \rho} e^{\zeta_2 T} \right) \left(1 + \frac{\rho}{\zeta_2} \right) \tag{A50}$$

Using the equilibrium condition found above, $\zeta_1 = \zeta_2 \equiv \zeta$, the last equation can now be reduced to:

$$e^{\zeta t} = e^{\zeta T} \frac{\rho}{\zeta + \rho} \tag{A51}$$

The right hand side is constant. The condition holds for all $0 \leq t \leq T$ if $\zeta \equiv (1-\pi)\mu - \rho = (1+\phi)\nu - \rho = 0$.

4. Banking in The Netherlands; a description of the data

4.1 Introduction

For many centuries technological change has had a profound impact on industrial production. The phenomenon that technical progress also affects the services sector is of a relatively recent date. Over the last decades it can be seen that, thanks to the development of computer and telecommunications technology, the service sectors have been able to broaden their scope of operations and their product range, and that marked rises in productivity have taken place. This has contributed to a gradual process of structural change in most western economies: a growth of services at the expense of industrial activity. Beside that, introduction of computer and telecommunication technologies, together called information technologies, has facilitated standardization and diversification of production. Mass production of standardized types of services for large markets has developed. Information technology also facilitates decentralization, a process which can have a profound impact on the division of labour, within and between organizations, and on patterns of employment, like part-time and self employment. Thus, in a number of respects the services sector has been affected by the emergence of information technology.

One of the sectors where these trends have been strongest is the banking industry. Computers have been introduced in banks on a massive scale at a relatively early date, and telecommunications networks have been set up in the recent past. These developments in information technology have led to a range of new banking products, an enormous rise in output and a spectacular drop in costs over the last decades. A set of data on banking activities in The Netherlands was available. For the reasons mentioned, it seemed to be of interest to use them for the empirical testing of the foregoing models. The first part of this chapter is devoted to a general description of banking in The Netherlands. The second part deals with a description of the data set that will be used for model analyses.

4.2 Banking in The Netherlands: an overview[1]

The banking and insurances sector is one of the larger sectors in the Dutch economy. In 1987, it produced one and a half times more gross value added than the largest industrial sector, and employed more people than any industrial sector, except for the sector of metal products and optical products. It produced in this year 6% of the total gross value added of the private sector and employed 4.5% of total labour in the private sector. For services as a whole, these figures are 57% and 60% respectively.

The banking industry is a conglomerate of many different financial services firms, most of which carry out several functions. According to the International Standard Industrial Classification (ICIC), the Dutch banking industry can be divided into four branches. The first branch consists of the Central Bank and general commercial banks (wholesale, retail, trading and investment banks); the second branch comprises cooperatively organized (agricultural) banks, postal giro services and savings banks; the third consists of other credit and financial institutions like building societies and brokers; finally, the fourth contains complementary financial firms like commission-agents in bonds and stocks and financial administrative firms. The larger part of the banking industry falls under the first two headings, general commercial banks, cooperatively organized banks, postal giro services and savings banks. In The Netherlands, these branches are presently dominated by six large banking organizations, accounting for at least between 80% and 90% of the banking industry's economic output and more than 90% of its employment.

The banking industry has four functions. Firstly, we mention its intermediating function in the payment system. Secondly, 'assets management' is important, by which banks are directing the composition of the assets side of the balance sheet. Activities rated among this function are participations and financing (loans, credits, mortgages and the like). The third function consists of activities directed towards the acquiring of financial means like savings and (demand and time) deposits, classified as 'liabilities management', indicating its effects on the liabilities side of the balance sheet. Finally, banks perform a number of other financial services like issuing shares and stock-jobbing on the one hand, and services originally not belonging to the banking profession, like acting as an agent on behalf of insurance companies and travel-agencies, on the other hand. Technological change in banking has mainly affected the first function mentioned, the intermediary function in the payment system. The six largest Dutch banks take care of the payment system in The Netherlands.

4.2.1 Main developments

Until the early sixties the banking world was relatively quiet. Traditionally, the different banks were strongly specialized and their activities were restricted to their own territory. The former postal giro services took care of the mass payment traffic of wage earners and consumers. The trading banks (general commercial banks) financed loans and credits for trade and manufacturing and accepted money deposits and savings from the same economic sectors as well as from the wealthy. Cooperatively organized agricultural banks financed activities and acquired savings in the agricultural sector while labourers had their accounts at a savings banks or with the postal giro service.

[1] This section and the next draw partly on the work of De Wit.

Since the mid-sixties, however, the situation changed dramatically. The prosperous economic growth of the whole economy led to a great demand for loans and credits. This resulted, on the one hand, in a series of mergers among the banks and, on the other hand, in the penetration of each other's markets through product differentiation. Wholesale banks started to operate on the retail market with the objective of acquiring savings to finance industrial loans and credits. Conversely, cooperatively organized banks with huge savings balances entered the wholesale market to sell loans and credits. Thus, beside the process of concentration, a process of branch blurring started and competition increased.

While the number of independent banks decreased dramatically, the number of offices in the country increased until the beginning of the eighties. Since then a reversal of the trend can be seen, a decrease of the number of offices. The total number of offices grew steadily from 7520 in 1971 to more than 8600 in 1981, and then decreased until 8232 in 1986. In 1986 the number of inhabitants per bank office can be estimated at 2616, or at 1765 if post offices are included. Internationally comparable figures are 2310 in the United States and 1524 in France.

The trends in the beginning of the 1990's are dominated by processes of reorganization and strategic reorientation. Partly this is happening in anticipation of the establishment of the single European market after 1992. Take-overs, mergers and strategic alliances across European borders enable banks to offer services on a European or world-wide scale, and to diversify and expand their range of financial products. This dynamism in the branch has lead to heightened competition and a pressing need to decrease costs. More attention than before is devoted to difficult matters of cost calculation and adequate pricing of different services. The restructuring of employment that has started in the eighties continues: more and more administrative jobs disappear and ever more commercial jobs are being created. These trends are visible in the local branch office networks of the banks. Mergers and the exploitation of economies of scale may well lead to the closing down of local offices. After a period of steady growth, employment is expected to decrease slightly in the early nineties.

4.2.2 Economic indicators

Table 1 gives an indication of the growth of banking in comparison to the market sector as a whole. It gives an overview of gross value-added in 1980 prices for the period 1971-1986.

Table 1: Economic growth of banking and the market sector at large (amounts in Dutch guilders $\times 10^9$; prices of 1980).

	banking industry			whole market sector		
year	value added	index number	annual rate	value added	index number	annual rate
71	6	100		226	100	
76	9	155	9.1%	268	119	3.5%
81	12	206	5.9%	287	127	1.4%
86	14	237	2.9%	307	136	1.4%
71/86			5.9%			2.1%

Sources: CBS, Statistical Yearbooks 1971-1987;
CPB, Central Economic Plans 1971-1987.

The figures indicate that the banking industry, compared to the whole market sector, flourished during that time period. Average annual growth rates in each period of five years were well above the comparable growth rates of the market sector at large, with an average annual growth rate over the last fifteen years of 5.9% versus 2.1% for the whole market sector.

As mentioned, the banking sector carries out different functions, the payment function, the 'assets' and the 'liabilities' management and other financial services. In Table 2 we present some indicators referring to the first three functions. The indicators do not represent value added but the aggregated balances figures at the end of the year. Although they therefore do not represent the development of economic output, they give us a rough idea of how total production volumes of different banking activities have been developing.

The average annual growth rate of payment transactions of more than 25% is striking. The highest growth rates, however, occurred in the beginning of the seventies (an average annual rate of 75%), while the growth rates during the last ten years varied between 5% and 6%. The banking industry has only been capable of processing these huge volumes because it changed to an automated payment system.

The other two functions have been developing more or less parallel to the banking industry as a whole, growing at about 6% per annum. Retail functions both at the 'assets' (consumers credit and mortgages) and at the 'liabilities' side (savings balances) have experienced lower growth rates than wholesale activities like loans and debtors balances at the 'assets' side and near money at the 'liabilities' side.

Table 2: Development of different banking functions (index numbers 1971=100; average annual growth rates 1971-1986).

Function		Payment	Assets Management				Liabilities Management	
Year	Value Added	Transactions	Loans	Mortgages	Cons. credit	Debt. balanc.	Savings balanc.	Near money
76	155	1680	167	205	195	211	107	225
81	206	2194	166	165	179	301	155	217
86	237	2945	223	201	177	248	160	306
71/86	5.9%	25.3%	5.5%	4.8%	3.9%	6.2%	3.2%	7.7%

Source: CBS, Statistical Yearbooks 1971-1987.

4.3 Technological developments

Technological developments in the banking industry started in the field of the payment system. One may distinguish between countries that might be characterized as having a 'cheque' payment system and others having a 'giro' payment system. Table 3 shows clearly that The Netherlands has developed a 'giro' payment system.

The banking industry's path of technological development, its technological trajectory, differs according to the prevailing payment system. Technological developments in countries like The Netherlands with a 'giro' payment system started with the automation of, firstly, records of the account of clients in the sixties and, secondly, of giro transfers. This was accomplished by investments in big mainframe computer systems in connection with the development and introduction of new 'regular' payment instruments like giro salary accounts and automated

Table 3: Composition of Payment Instruments in 1983

	Transactions		Amounts	
	Cheques/ Creditcard	Giro Transfers	Cheques/ Creditcard	Giro Transfers
France	85	15	10	90
Netherlands	23	77	1	99
United Kingdom	71	29	10	90
United States	97	3	28	72
Germany, F.R.	11	89	16	84

Source: Bank for International Settlements (1985).

debits and credits. In 'cheque' countries, however, one was less able to develop comparable automated instruments for periodical payments like salaries, mortgages payments and monthly rents. Consequently, payers were forced to continue writing cheques periodically. Technological developments in countries like the USA, the UK and France were mainly directed at the automation of the labour intensive processing of cheques.

The automation of payment transactions began after World War II, when the economic revival induced a strong growth of the payment traffic. Mechanical bookkeeping machines were introduced in the fifties. The first mainframe computer, combined with the punched card, made it possible in the beginning of the sixties to automate financial mutations in the account records of the administrations at the central offices of the banks. This process of central automation was reinforced by the establishment of a so-called automated clearing house. The decentralization of computer and communication technology started with the automation of local offices and banks and was intensified by the establishment of information networks. Consequently, the following phases of technological development can be distinguished:

1. Central automation
   a. at the head offices of individual banks (1960-1970);
   b. of clearing houses and external integration with business clients (1970-1980);

2. Decentralized automation
   a. of local branches' back offices and front offices (1980-1990);
   b. of information networks as the merger of computer and communication technologies (1990-..).

Both stages of central automation took place at only a few production centres of commercial banks and clearing centres. Its applications concern the centralized registration of the loans, savings and securities accounts, but above all the central administration of current and salary accounts, as well as the payment transactions involving debiting and crediting these accounts. The effect of this process of technological development mainly was a strong increase in the labour productivity associated with the central processing of the payment transactions.

We now witness, in the phases of decentralized automation, as opposed to the period of central automation, a process of technological development in which certain offices of local banks could be qualified as early adopters and others as laggards regarding the adoption of computers and communication technology. In other words, the process of diffusion of technologies through the branch network of banks is paramount. Moreover, the effects on employment will not be restricted to a small group of production workers at clearing centres and production centres at the head offices, but will apply to the majority of bank employees, about two-thirds of which are working at local banks' offices throughout the country.

4.3.1 Local banks and offices

The technological development at branches and local banks can be divided into five stages; the first two interact with the phases 1(a) and 1(b) of central automation and the last three are a further differentiation of phase 2(a) and phase 2(b) as distinguished above.

1a	Manual production process making use of mechanical bookkeeping machines (1960-1970);
1b	Optical character recognition (OCR) equipment (1970-1980);
2a(i)	Mini-computer system with back-office terminals (1980-1985);
2a(ii)	Counter terminals in front office (1980-1990);
2b	Installation of ATM's (1985-..).

During phase 1 of central automation, until the early seventies, the technological developments at local banks and offices were restricted to the use of mechanical bookkeeping machines. From the beginning of the seventies onwards, OCR equipment came into use. The equipment was used for typing the payment orders of clients on 'counting-slips' which could be read optically by the mainframe computers at the central head offices of the banks. Most machines were equipped with controlling functions correcting simple mistakes and producing a cleaner input at the mainframes. Up to this stage there were no major differences regarding technological developments at the local offices between the organizations involved.

The start of computer and communication systems at local banks and offices, stage 2a(i), can be dated around the middle to late seventies. This led to a productivity increase in back-office work. This had little consequences for employment, however, due to an increase in output. The next stage of technological development at local banks was the installation of terminals at the bank's counter and so-called quick-cash terminals, stage 2a(ii). These counter terminals are 'smarter', having more in-built functions. The possible effects on employment and the organization of work are more pronounced. In particular, much of the work of cashiers and counter clerks is now being automated. Since the end of the eighties, banks are installing cash dispensers or automatic teller machines (ATM's) on a large scale, stage 2(b), thereby displacing even larger parts of labour at the counter. These systems require not only automated account management, but also a communication network between local offices and the central computer system.

The adoption of more advanced information-technology-based production techniques has consequences for commercial policy: before the introduction of off-line counter terminals the commercial policy could be qualified as product oriented, implying that employees were specialized in certain products, for instance, cash payments, insurances, consumer credits or business loans. With the introduction of counter terminals in the eighties, commercial policy changed into some type of 'integrated client management'. This kind of management implied that clients were in principle served the whole range of banking products by a 'personal banker' who was assisted by some counter clerks. This meant on average that the commercial employees were expected to have a higher education and more skills, but at the same time that they also had to fulfil routine activities and tasks in which they were not so well trained.

Banks are currently in a stage of establishing, implementing and expanding information networks, stage 2b. An information network originates when the management at offices and local banks can retrieve data (about accounts and characteristics of clients as well as external financial data) from the central or local computer systems, use the data with the aim of gathering information for management purposes and possibly send the newly processed information back to the central computer systems or to clients. The actual applications of the networks are considered to mature in the nineties.

4.3.2 Banking product and process development

In banking, it is not so easy to distinguish between process innovation, on the one hand, and product innovation, on the other. Usually, the two develop hand in hand. For instance, the process of central automation at head offices of banks and clearing centres could not have been realized without a simultaneous introduction of new products like the salary and current accounts, automated debits and credits, and giro-cards inviting payment. The fast increase in the number of machine readable payment instructions at the automated clearing house (from 5.8 million in 1970 until 278 million in 1985) illustrates the growth of the bank's new payment products. Diffusion of new process technology is reflected in the gradual introduction of new product innovations.

Asset management is relatively unaffected by information technology so far. The contracting of loans still predominantly consists of face-to-face negotiations and an analysis of the client's financial data. The use of personal computers for financial analysis, the recording of the client's financial details, the word-processing of standard financial contracts as well as the associated administration, have only just started. Applications in the field of mortgages are a little bit more advanced, in the sense that it is possible to produce standard offers; however, it is felt that the applications do not go much further than a qualitative support of the negotiations with the client. In the case of insurances and travelling, some use is made of on-line communication when processing the contract for an insurance policy or the booking of a journey.

Finally, some remarks should be made regarding cash dispensers, point-of-sale and electronic banking terminals. These products were mainly developed in the United States. In Europe, 'cheque' countries like the United Kingdom and France followed soon. The developments in 'giro' countries like The Netherlands and the Federal Republic of Germany started much later, because originally no action against unsecured cheques was needed. Nowadays high labour costs of teller transactions and pressures from retailers and petrol companies accelerate the diffusion of cash dispensers.

4.4 The data set

Some aspects of the developments described in the foregoing sections of this chapter will be explored quantitatively further on. This will be done using a data set collected by a large cooperative banking organization, with over 900 affiliated banks. These affiliated banks are legally independent, but cooperate in a larger organizational structure. With regards to investments in new technology, they are off course restricted by their membership of the cooperative in their decisions on the systems and the standards for which they opt. Nevertheless, they have considerable autonomy in deciding on the speed of their technological progress: they can choose to adopt early or late. There may have been some centrally imposed constraints also with regards to the speed of adoption, e.g. an ultimate day of compliance to certain standards, but information on this is not available. From these 900-plus banks, a representative sample of about one hundred banks is engaged in a yearly internal efficiency monitoring project. The earliest data from 1979 are rather sparse, but as time proceeds available information gets more detailed. The last observations available were from 1987. Unfortunately, the registered sample changes somewhat every year and the system of registration of data has been revised in 1985, making the data set a bit fragmentary and not fully consistent. Nevertheless, information is rather detailed and a closer look is warranted.

There are a number of data types that are contained in different data sources:

1. Automation data: investments in six types of automation equipment, of 119 banks, over the years 1979 to 1987.
2. Cost data: costs per bank, per year (1985, 1986, 1987), per product, per cost type, for close to one hundred banks per year.
3. Employment data: detailed data on employees per bank, concerning profession, salaries and education.
4. General data: about 30 figures per bank on production volumes, aggregate costs, depreciations and the like, for 1984 to 1987, for the same banks as mentioned under 2.

We shall consider data on automation, on the level of activities and on costs in some detail.

4.4.1 Automation data

These data give a quantitative and a qualitative indication of the spread of information technology among local banks. Not only the amounts spent on automation equipment, but also the types of machinery bought, are registered. There are figures on investments in six types of automation equipment, by 119 banks, over a period of eight years. The six categories of investments are: 1) front office automation equipment; 2) either back or front office automation equipment (undifferentiated);[2] 3) personal computers and networks; 4) rapid cash registers; 5) automatic teller machines; 6) a system for travel bookings. From these data we can make inferences about the technological level of a bank.

Investments in automation equipment on the local bank level have increased considerably over the last decade. Looking at investment streams in automation over the period 1979 to 1987 (Figures 1 and 2), we notice that there has been a marked upsurge in 1984 and 1985. The bulk of these investments went into front office automation programs. Since then there has been a slight decline in automation expenditures. Investments in personal computers is a relatively constant stream over this period. Investments in rapid cash terminals have peaked around 1985, and investments in automatic teller machines are taking off towards the end of the period under consideration.

[2] There is equipment that can be used both by banks that have an automated front office and by banks that only have an automated back office. From investments in these type of machines no inferences about the precise technical level of the bank can be made. This type of equipment is contained in the second category. The first category contains all equipment that can only be used to automate the front office.

Figure 1

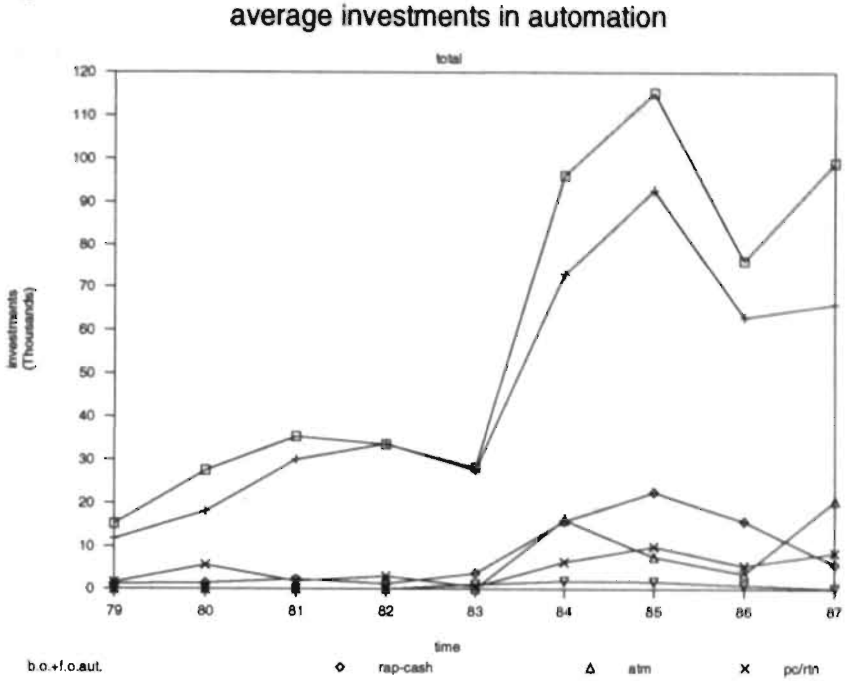
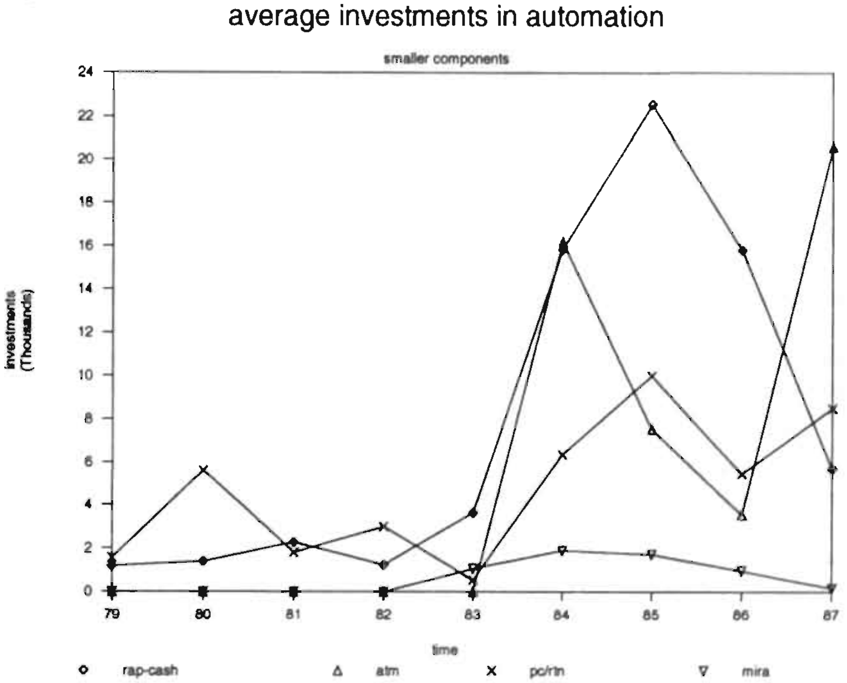
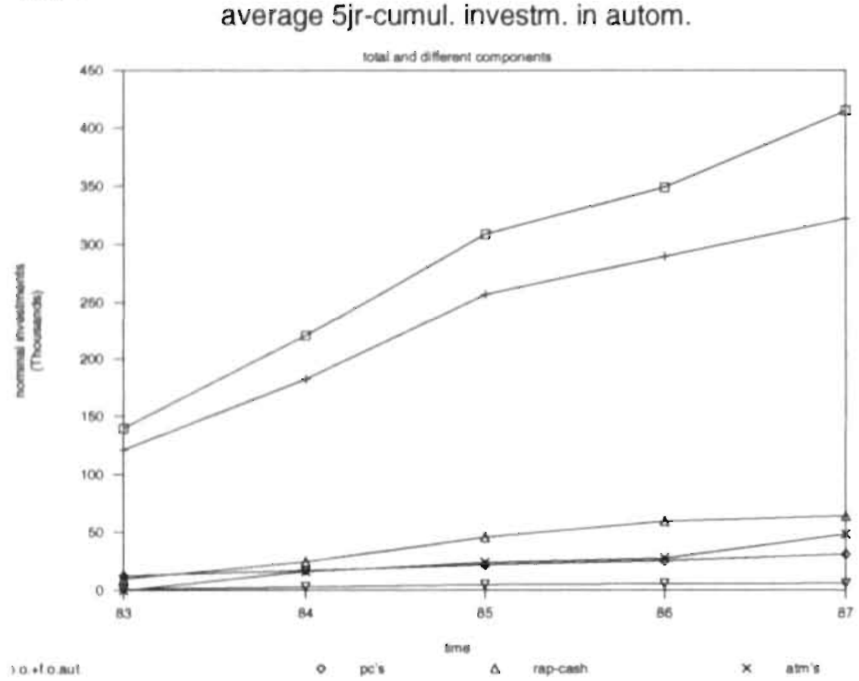


Figure 2



Investments in automation equipment are generally depreciated over a five year period. Figure 3 shows average nominal investment, cumulated over five years, between 1983 and 1987. This gives an impression of the average rise in the size of the stock of automation equipment in banks. Over the years 1983 to 1987 the average total stock has almost tripled, indicating that the average bank indeed witnessed some major expansion in this respect during our sampling period.

Figure 3

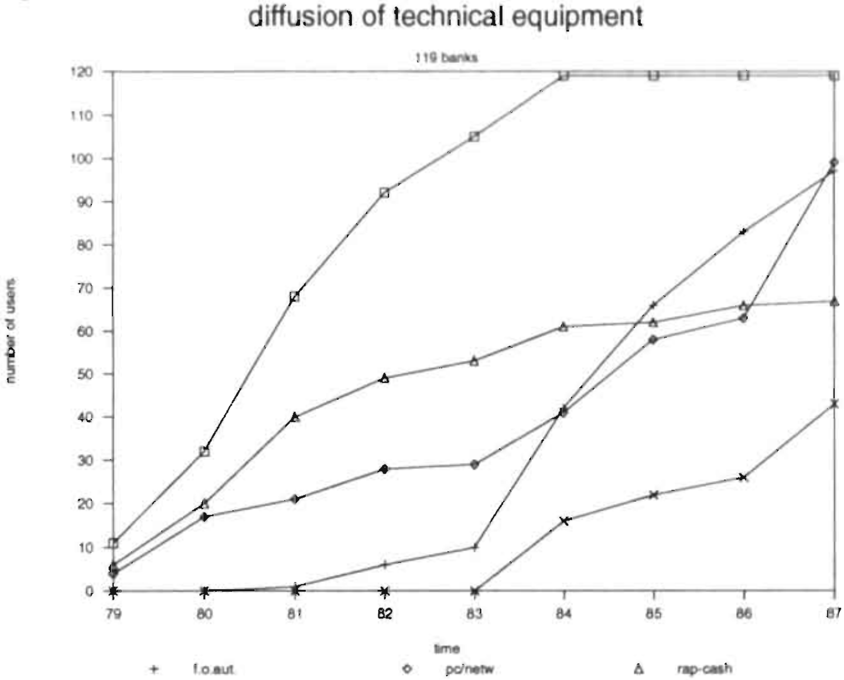


Most banks have installed back office automation equipment at a rather early stage. In most cases, front office automation has equipment been added to this a few years later. Front office automation is thus an extension, not a replacement, of back office automation. Automation of the front office demands relatively large investments and usually requires a far-reaching restructuring of operations and training of personnel. Although back office automation and front office automation are not two distinct 'techniques' in banking, in the sense that the latter replaces the former, one can say that a bank having an automated front office is on a higher technological level than a bank with only back office automation.

In our data set it is not indicated explicitly at what moment the front office has been automated. However, there are some systems that are characteristic and necessary for front office automation, and we assume that from the moment onward that a bank has invested in such a system, it has an automated front office at its disposal. However, the bank's expenses on office automation systems from that moment onward do not have to be restricted to front office equipment, but can also include further investments in automating the back office. Conversely, we assume that as long as a bank has not invested in the characteristic front office automation machinery, its automation expenditures that do not fall in any of the other categories are for the automation of the back office.

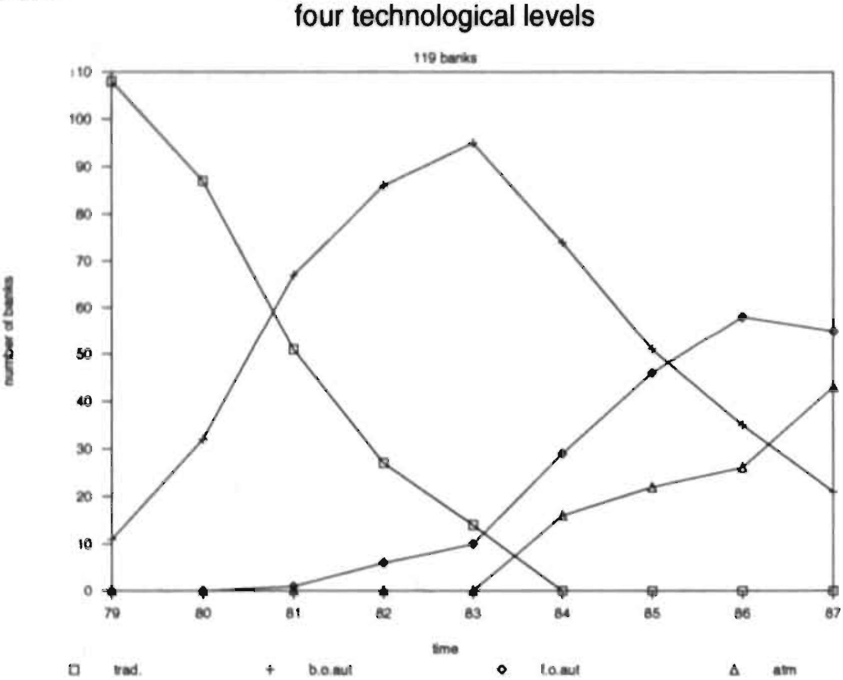
Using the dates of the first investments of a bank in a certain category of equipment, we can construct a graph that illustrates the diffusion of types of automation equipment, and thus of certain production techniques. Figure 4 shows for eight years how many banks from the sample of 119 possessed a certain system. The curves indicate that both back and front office automation diffused over a period of approximately five years. The use of personal computers and networks diffused slower. The diffusion of rapid cash terminals slowed down after the introduction of front office automation and automatic teller machines. The diffusion speed of ATM's increases after 1986. The curves follow the sigmoid shape familiar from diffusion literature: the rate of adoption of systems at first increases and later decreases. The similarity in shape and slope of the curve of back office automation and the curve of front office automation is remarkable.

Figure 4



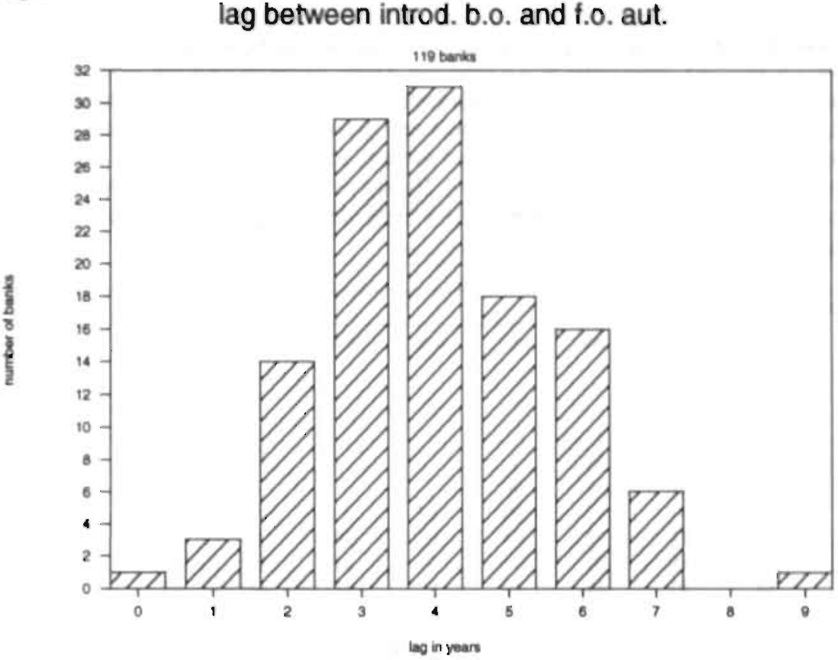
As mentioned, it is more correct to think of different banks as operating on a different technological level, rather than as operating distinct techniques. Thus we distinguish at the local branch office between the first level, on which banking operations are still not automated, the second level, on which only the back office is automated, and the third level, on which both the back and the front office are automated. Investments in back and front office automation constitute by far the largest components of the automation budget, but there are data on a number of smaller expenses available. A closer look reveals that, as far as the timing is concerned, back office automation investments precede all other expenses. Banks automate the back office before investing in front office equipment, rapid cash terminals, personal computers and network facilities. Investments in automatic teller machines in general follow as the latest developments. The coming to maturity the last category falls outside the scope of the data set, but here also major investments in machinery and network connections are required. The cash dispenser is a more advanced alternative for many counter operations and for the rapid cash terminal. Therefore we distinguish a fourth level of technological progress, for which the automatic teller machine is indicative. Figure 5 shows over a period of 9 years the distribution of 119 banks over these four technological levels.

Figure 5



Considering the similarity of the diffusion patterns of back office automation and front office automation, one may wonder whether the banks that are first to automate the back office are also the first to automate the front office. Figure 6, presenting the number of years between the introduction of back office automation and of front office automation equipment, reveals that this is generally not the case. It seems that among the banks in the sample there is not a clear division between first movers and laggards.

Figure 6



4.4.2 Firm size data

Beside data on investments in automation equipment, there are two other important categories of information available: data on activity levels and data on costs. Together these two categories yield a picture of scale, inputs and outputs of a bank. In this section we shall deal with data on levels of activity, in the next with data on costs.

Differences in the scale of operations or in size of a bank can be factors that lead to differences in technical capabilities, in the ability to exploit economies of scale, the ability to finance large investments or the willingness to run financial risks. These differences may cause differential speeds of technical progress or of times of adoption of new technology. Scale of operation can be measured by looking at inputs, at outputs or both. Given that we are after the relationship between inputs, outputs and technological progress, the main inputs of interest are labour and investments in high technology equipment. The main outputs are the outputs produced using information technology, which are connected mainly with maintenance and monitoring of accounts, transfers of payments, and the like. For 1987, we have information on these variables for 82 banks.

On the output side, the following categories are distinguished: 1) current accounts, 2) savings accounts, 3) loans, and 4) mortgages. Concerning these four output categories several data are registered per bank:

1. The number of current accounts, savings accounts loans and mortgages;
2. The total value on savings accounts, in loans and in mortgages;
3. The number of mutations in current accounts, saving accounts, loans and mortgages.

Of the total number of accounts, 62.5% are savings accounts, 26% are current private accounts without credit, 5.8% are loans, and 5.6% are current and private accounts with credit. Of the total number of registered mutations in current and savings accounts and loans, 92.5% are in current accounts, 5.5% are in savings accounts and 2% are in loans. The ratio of the value in loans and the value in savings is 1.3. In Table 4 below these size indicators are listed with their mutual correlation coefficients.

Table 4: Correlation coefficients size indicators

	mca	awc	asc	msa	nsa	asa	mln	nln	aln	nmo
mca	1									
awc	.73	1								
asc	.91	.58	1							
msa	.95	.67	.95	1						
nsa	.94	.68	.96	.98	1					
asa	.89	.65	.91	.92	.96	1				
mln	.81	.63	.90	.87	.89	.87	1			
nln	.84	.61	.93	.92	.94	.93	.96	1		
aln	.87	.67	.85	.89	.91	.93	.81	.89	1	
nmo	.52	.34	.40	.49	.49	.55	.41	.48	.62	1
amo	.48	.31	.35	.43	.43	.50	.33	.41	.59	.98

Legenda: mca: number of mutations of current account; awc: accounts with credit; asc: accounts without credit; msa: mutations savings accounts; nsa: number of savings accounts; asa: amounts on savings accounts; mln: mutations in loans; nln: number of loans; aln: amounts in loans; nmo: number of mortgages; amo: amounts in mortgages.

Table 4 shows that current account activities and savings activities are highly correlated, but that those two are less correlated with loan activities and even less so with mortgages. Mutations in current accounts, numbers of current and private accounts without credit, numbers and mutations of savings accounts, and amounts on savings accounts are all highly correlated. The number of loans is highly correlated with savings activities and the number of accounts without credit. The amount of loans is highly correlated with the amount of savings. Only mortgage activity seems to be rather independent of the other variables. Thus, Table 4 indicates that a number of variables can serve well to represent a sort of general output measure of the activities of the bank. An aggregate output measure will be rather robust against changes in the weights of the components.

Concerning input data we have to rely mostly on cost data, except for data on labour inputs, which are in volume terms (number of employees weighted by numbers of days worked per year). Table 5 shows that labour volumes are highly correlated with output. This suggests that at the aggregate level labour productivity is rather stable, despite differences in technology.

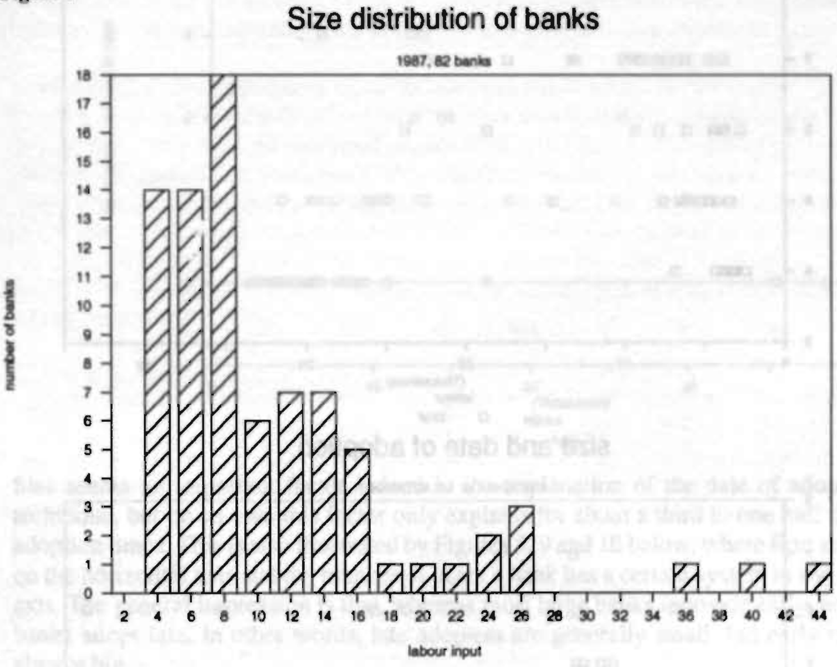
Table 5: Correlation coefficients labour and size, different indicators

	mca	awc	asc	msa	nsa	asa	mln	nln	aln	nmo	amo	lab
lab	.92	.71	.90	.93	.95	.94	.81	.88	.96	.65	.62	
lad	.92	.72	.89	.93	.94	.94	.83	.89	.96	.63	.60	.98

Legenda: lab: total labour; lad: administrative labour; others: see under Table 4.

Labour is highly correlated with current account and savings variables, and less with loans and mortgages. Assuming that the size of the workforce is a good measure for the size of the bank, we can plot a distribution over bank-sizes (see Figure 7). The distribution is heavily skewed to the left: there are a lot of small banks, but only a few large ones. Plotting a distribution using other size indicators gives a similar impression (see Figures 11 and 12 in the appendix to this chapter).

Figure 7



Regression of adoption dates on size indicators shows that on average large firms adopt new technology earlier. More specifically, regression of 'years of back office automation' on the logarithm of mutations in current accounts and number of workers gives the following result:

dependent:	independent:	coeff.	t-stat.	r-sq.
years of b.o. autom.	ln(mut. curr. acc.)	1.605443	9.14	0.51
	ln(labour)	1.606521	8.72	0.49

Regression of 'years of front office automation in 1988' on the logarithm of changes in current accounts and number of workers yields:

dependent:	independent:	coeff.	t-stat.	r-sq.
years of f.o. autom.	ln(mut. curr. acc.)	1.694265	7.39	0.41
	ln(labour)	1.735044	7.38	0.41

Figure 8

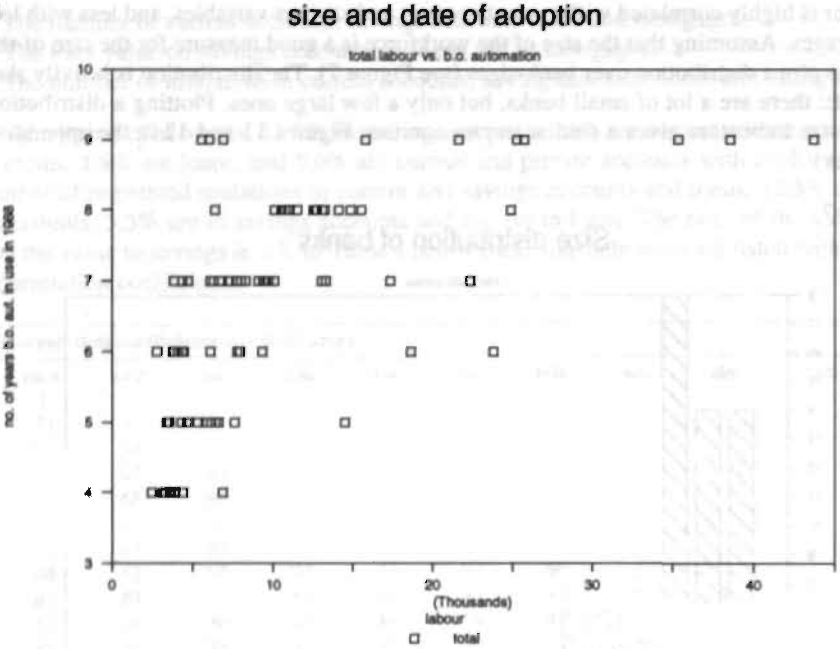


Figure 9

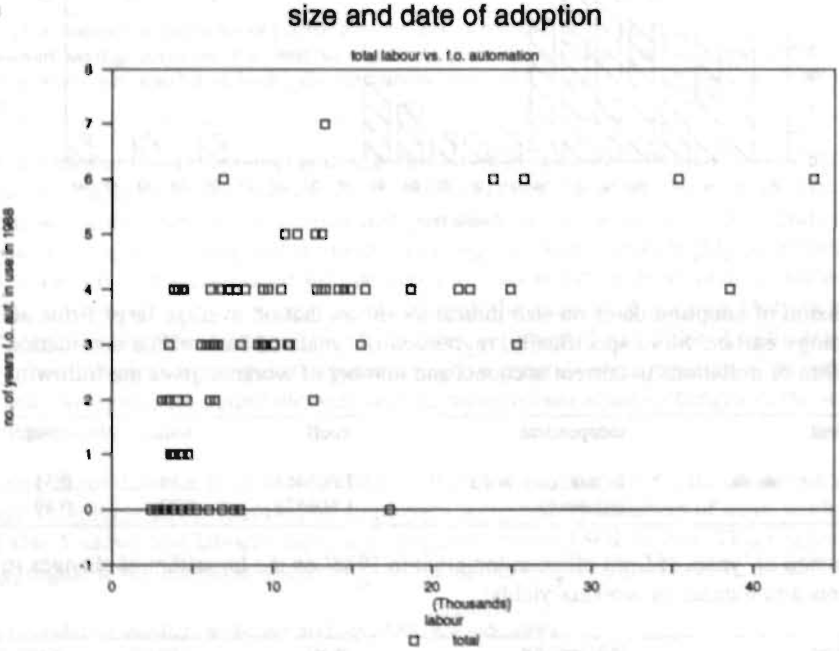
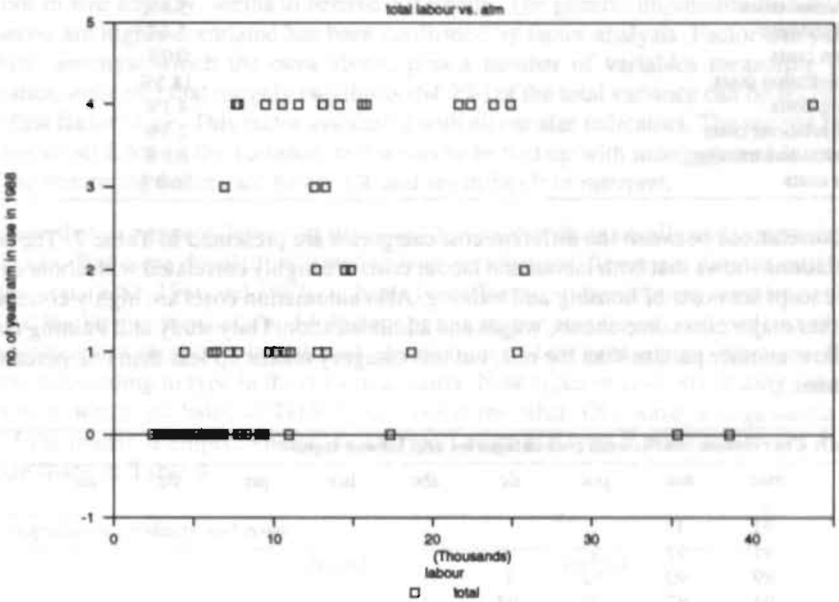


Figure 10 size and date of adoption



Size seems an important factor indeed in the explanation of the date of adoption of a new technique, but on its own this factor only explains for about a third to one half the variation in adoption times. This is also illustrated by Figures 8, 9 and 10 below, where firm size is displayed on the horizontal axis and the number of years a bank has a certain system in use on the vertical axis. The general impression is that, whereas most large banks innovate early, not all the small banks adopt late. In other words, late adopters are generally small, but early movers are not always big.

4.4.3 Costdata

A problem in analyzing data sets that are as differentiated as these, both in types of inputs and in types of outputs, is to distinguish between the various parallel production processes that link the different outputs to their specific inputs. A bank produces a number of outputs with the same inputs, making it difficult to assign costs to products. At the construction of the present database, attempts have been made to tackle this problem at the source and to register inputs together with their destination. Therefore the cost information has a rather elaborate structure.

Nine broad but rather unequal categories of costs are distinguished in our data set. More than half of the costs are wage costs. The other half is divided into eight cost types. Administration and office costs constitute about one sixth of the total; automation costs are only about 6% on average; the other categories are rather minor. Their percentages in total costs are as follows:

Table 11: Distribution of costs I over products

type I; in %	payments	saving	loans	total
management costs	0	11	10	6
administration costs	23	28	6	18
labour costs	53	38	71	56
depreciation	8	6	0	7
interest losses	8	6	0	7
other costs	8	11	13	6

Payment transfer operations are still relatively labour intensive, despite automation. Depreciation of equipment is a minor cost factor, only one twelfth of the total. Clearly, loans are the most labour intensive products of the three.

Table 12: Distribution of costs II over products

type II; in %	payments	savings	loans	total
commercial costs	3	12	67	25
cashier and counter costs	30	25	0	21
processing costs	36	25	3	22
indirect costs	31	38	30	32

Indirect costs account for about one third of the costs of all these products. Important differences are in commercial costs, which are important mainly in financing operations, and in cash, counter and processing costs, which are the bulk of costs in payment transfers and, to a lesser extent however, in savings.

4.4.4 Scale economies

Given these detailed data on costs and our measures for firm size, one can try to determine whether scale economies are important to the banks in the data sample. There are economies of scale, if an x% increase in size leads to a less than x% increase in costs, and there are diseconomies of scale if the increase in costs exceeds x%. To determine this size elasticity of costs, the logarithm of costs has been regressed on the logarithm of size, plus a constant. For size two proxies have been selected: number of mutations in current accounts and total labour input. For costs 14 measures have been tried: total costs, a subdivision of total costs in 9 type I items, and another subdivision of total costs in 4 type II components. The results are displayed in the following table. Only estimated elasticities are given. T-statistics indicate that all estimated coefficients are significant and that mostly the constant, which is sometimes positive sometimes negative, is also significant. The first two columns give results of estimations with all banks in the sample; in the last two columns the banks without front office automation have been left out, to get a more homogeneous sample.

In Table 13 the estimated labour input elasticities of different types of costs, the second column, are given in ascending order per sub-category. If we would order using the first column as key, the table would change only marginally. Thus, as far as the ordering is concerned, the two measures of size yield approximately the same results. However, the elasticities in terms of mutations of current accounts are generally lower than in terms of labour inputs. Scale economies are more pronounced if we look at this highly standardized output than if we look at labour in general.

Table 13: Results regression of costs on size (in logarithms)

cost type	mut.c.a. all banks	labour all banks	mut.c.a. f.o.aut.	labour f.o.aut.
management costs	0.66*	0.71*	0.64*	0.70*
public relations costs	0.69*	0.74*	0.64*	0.70*
study and training costs	0.72*	0.79*	0.69*	0.78*
automation costs	0.90*	0.92*	0.85*	0.89*
postal services costs	0.95*	0.97	0.94*	0.98
administration and bureau costs	0.98	1.01	0.96	1.01
labour costs	0.97	1.02*	0.96	1.04*
housing costs	0.95	1.03	0.94	1.04
telecom costs	0.97	1.04	0.97	1.05
processing costs	0.88*	0.90*	0.87*	0.91*
indirect costs	0.86*	0.92*	0.84*	0.92*
commercial costs	0.96	1.03	0.94	1.03
cashier and counter costs	1.11*	1.15*	1.10*	1.16*
total costs	0.94*	0.99	0.92*	0.99

Coefficients that are significantly different from 1 at the 5% level are indicated by a *.

The coefficients obtained for total costs indicate that there are slight overall economies of scale. Scale economies are most pronounced in costs of management, public relations and study and training, but also in automation costs. Remarkably there are virtually no scale economies observed in costs of administration. Scale economies can neither be shown in overall labour costs. However, if we consider components of type II which consist to a large extent of labour, there seem to be counteracting movements which balance out. Larger banks seem to be able to save on processing and indirect costs, but spend extra on cashier and counter work. Different reasons could be advanced for this last finding: larger banks offer more products at their counters; customers with more complicated demands go to larger banks; counter work is a category in the time administration in which everything not administered otherwise ends up, and in large banks there is more undefined use of time and structural overcapacity.

Furthermore, it is remarkable that there are scale economies in training costs, but diseconomies in labour costs. An explanation for this finding could be that larger banks have on average more expensive employees, e.g. more experts, but spend less on training them. The the second column, in which labour inputs are used as size indicator, shows diseconomies of scale if we look at costs of administration, of labour, housing, telecommunication, commercial activities and counter services. This also seems to indicate that as banks get larger, they employ relatively more expensive personnel, although less management personnel, and put more work into counter service and commercial service.

It might be that what appear to be scale economies above, are in fact the effects of the earlier adoption of new technology by larger banks: larger banks produce more efficiently, because they apply more efficient techniques. To account for this effect, the equations have been re-estimated, leaving out the banks that have not installed front office automation equipment yet. From the table it is clear that the earlier impression of scale effects is reinforced, rather than diminished. This is especially apparent, if we consider the results obtained with mutations in current accounts (column 3). All coefficients here are below the overall estimates. It thus seems that new technology compensates for diseconomies of scale in the earlier technique. The coefficients obtained from equations with labour inputs as the dependent variable point in the same direction, although the evidence is less decisive.

The overall picture points out that there are certain economies of scale, especially in the production of standardized outputs. These scale economies are obtained through lower processing costs. Scale economies, however, are hardly translated into lower overall costs, because the extra margins seem to be invested in personnel doing commercial work and especially delivering counter services.

4.5 Selection of data for model testing

The data presented above will be used in the next chapter to test some aspects of the models in chapter 3. These models of the firm mainly relate three key variables to each other: output, variable factors of production and investment. The model is built on the assumption that output is sold on the market against the current market price, that variable inputs are rented on a short term basis and involve no long term commitments, and that investments are long term commitments, aimed at expanding production capacity and at reducing the need for variable inputs. The effectiveness of investment might be dependent on the technological level of the firm. In this section we shall attempt to find reasonable data from the data set, as proxies for these variables in the model.

First of all, consider output capacity. In the model, output is assumed to be homogeneous and of constant quality over time. The output of a bank, however, is far from homogeneous. One can look at a bank's output from different perspectives. Above we considered the functions of the bank: it acts as intermediary in payment transactions, as creditor, as debtor, as financial adviser and so on. Thus the bank guards deposits, grants credits and loans, gives advise, and transfers money. One can also look at the product of banking in a more 'physical' sense of the word. Seen from this perspective, the product of a bank is largely administrative output: the bank produces accounts, updates on account figures, mutations in accounts and transfers of money, calculations of interest, things in your mailbox. This type of administrative output is produced in a highly standardized way, mainly by a labour force consisting of cashiers and counter personnel, administrative personnel in the back office and data typists. Production of this type of banking product is progressively automated by the introduction of back office automation, front office automation, automatic teller machines, optical character recognition equipment, telebanking, etc. Beside this rather physical mass product, the bank produces also less standardized output, like business credits, mortgages and financial advice. These products are more customer tailored and more labour intensive in production. Here service plays a larger role and more expertise is required. Labour generates more value added here than in the production of the standardized output. These latter activities can only be automated to a very limited extent.

The model we to want test describes the development of a firm producing a standardized type of output, which is produced with limited fixed and lots of variable factors of production, but which could be produced more efficiently if there were investment in technical progress. If this model is to describe the activities of a bank, we should disregard the output of financial advice, credit grants and loans. What the model might describe accurately, is the development of the production of the administrative product. Margins on this product are low, volumes are large and have been growing over the last decades. There is ample room for automation by introducing computer and telecommunication technologies. The introduction of these technologies is a lengthy process of experimenting, learning, improving and expanding. Computer equipment replaces low qualified labour which gets progressively more expensive due to rising labour costs. This type of labour is a relatively variable factor, because there is a fairly large natural turnover and no large investments in human capital have to be depreciated.[3] All this accords well with the assumptions of the model.

There are two categories of data series in the data set that could serve as indicators for the volume of a bank's production of this standard output: there is information on payment traffic and there is information on savings activities. These two categories capture most of the standardised business of banks. Payment traffic is clearly the more voluminous output of the two. Looking at total costs, almost one half is attributable to payment traffic and about one seventh to savings (see Table 14). Between one third and half of the costs of labour are attributable to payment traffic and around one twelfth to savings.

Table 14: Costs of payments and savings as share in total costs and labour costs

percentage costs		payment traffic	savings
as share of total costs	1985	43.1	15.7
	1986	42.8	14.1
	1987	45.6	12.9
as share of labour costs	1985	44.1	8.9
	1986	41.2	9.1
	1987	43.0	8.6

The level of activity connected with payment traffic has developed differently from the level of activity concerned with savings. Whereas the former shows an increase in virtually every bank, measuring activity both in terms of numbers of accounts and in terms of account mutations, the latter shows a slight drop in about half of the total number of banks in the sample, going from 1985 to 1986. Going from 1986 to 1987, the number of savings accounts drops in almost every bank in the sample, but the number of mutations decreases in 12% of the banks. The model of chapter 3 describes growth of production, due to capacity increase, given an expanding or stable market. In its basic form, it does not describe contraction. It seems that in the period under consideration, the savings product of the bank was under pressure, probably due to factors outside the realm of elements which are accounted for in the model, like changes in savings conditions or interest rates offered. Because, on the one hand, payment traffic is a more important product of the bank, in volume and cost terms, than savings, and on the other hand, payment traffic activity grows over the period we consider, whereas savings activity declines, it seems most appropriate to apply the model to data on payment activities.

Information on payment traffic consists of data on the total number of current accounts and the number of mutations of accounts. Numbers of accounts give an impression of the size of a bank, maybe of the extent of fixed activities connected with monitoring. Mutations in accounts give an impression of the fluctuating activity of the bank. Since it is the actual production activity

[3] In the bank for which we have data, the turnover rates in the latter part of the seventies were above 12% and in first half of the eighties between 7% and 11%. The average job tenure of administrative personnel, desk clerks, data typists and secretaries was around five to seven years.

of the bank that we want to capture, it seems that mutations in accounts, which is a flow variable, would be a more appropriate series to use as proxy for production volume than numbers of accounts, which has more the feature of a stock variable. On the other hand, numbers of accounts may give more an impression of the capacity to produce than mutations. One option to construct a measure for output would be to use a weighted sum of the two series. Because no clue as to the right relative weights of mutations and of numbers of accounts is available, both series will not be aggregated but will be used separately. The number of mutations in current accounts of bank j will henceforth be indicated by Y_{1_j} and the total number of current accounts registered with bank j will be indicated as Y_{2_j}.

It should be kept in mind that, although any of the indicators proposed might approximate output satisfactorily, it is possible that none correlates highly with production *capacity*. The model describes the development of output capacity, which might develop differently from output itself, if there are large variations in the rates of capacity utilization. By taking output data as proxies for capacity, we assume that there is a stable relationship between capacity and production. If the firm raises capacity, it is assumed to be able to sell its additional output. Demand is assumed to be more or less price elastic, but markets are not saturated. Possible fluctuations in capacity utilization rates are not accounted for in the model. In the particular case of banking operations, this assumption is questionable: banks do not charge explicit prices for mutations of accounts. Therefore, the decision on the part of the customer to use this service of the bank is based on different arguments than the price or cost of transfer of money. Accounts are changed whenever there are payment obligations, need for cash, interest differences, and so on, but demand can hardly be assumed a function of price.

Secondly, consider variable factors of production. This is a variable for which it is even more difficult then for production capacity to select a data series to represent it. The problem is that almost any factor is variable at the margin, but, considering time periods of one year, is to a large extent fixed. A bank can decrease the necessary input of counter work or data input work, of mailing or paper or housing or cleaning, of management or administration. This it can do by investing in an increase of factor productivity. However, it is impossible to do away with counter work, data input work or management all together. Thus, in every category of factors that is registered, we can expect that there is some part expendable and a large part fixed. This means that to take any series as data for variable inputs introduces a measuring error, and to leave out any series introduces an error as well.

Nevertheless, it can be assumed that some inputs are more variable than others. Labour could be more variable than other inputs. Although labour contracts do not allow much flexibility on a short term basis, a high job turnover rate introduces some dynamics. Highly qualified personnel, like experts, managers and board, can be considered less variable than low qualified labour, like data typists and desk clerks. The higher job turnover rates of the last group are an indication which support this intuition. The low qualified labour produces the standardized output of the bank. Thus, if we model the production of a standardized banking product, there is something to be said for taking specific types of labour as a proxy for variable factors: administrative work, data input work, cashier and counter work. An alternative would be to consider total labour input. Although specific types of labour, say administrative labour, fits the concept of variable input more than total labour, it is also more prone to measurement errors and flaws in the registration. Therefore we use three alternative data sets for the estimations below. By V_{0_j} we designate the total labour employed by bank j. This has been measured by dividing total labour costs in a specific year by the appropriate price index. Furthermore, by V_{1_j} we designate total labour, used by bank j as input to the production of payment traffic, and by V_{2_j} data input work, cashiers work and counter work used for payment traffic. Thus V_{2_j} is equal to V_{1_j} minus the categories commercial and indirect costs of payment transfers. V_{2_j} is about 70% of total the data processing, cashier's and counter work.

On the one hand, there are data in the data set on numbers of employees and, on the other hand, there are data on costs of certain types of labour for the production of certain types of output. Since most employees perform different tasks in the organization, and our interest is in specific types of work as inputs to the production process, we take data from the files of costs of certain types of labour inputs.

Next, consider the choice of a series for the investment in progress. This seems relatively straightforward: investment in automation equipment is explicitly aimed at raising the productivity of the variable factors of production, and at raising production capacity. Most automation equipment is installed for the facilitation of payment traffic. A difficulty is presented by the fact that investments in automation tend to have a lumpy character. These investments tend to take place at large and irregular intervals and do not yield a smooth data series. One consequence of these bulky investments is that utilization rates will not be stable. Right after the completion of the investment project there will be overcapacity, which will be filled up gradually as the firm expands output. Another consequence of lumpy investments is that it takes a large effort of the organization to accommodate to the new way of producing. There is a considerable learning period, before the investment results in the aimed productivity rises. For modelling, this means that, first of all, production and capacity tend to diverge, and therefore production might not be a good proxy for capacity, and, secondly, that an investment at some period will have delayed effects on productivity in later periods. To account for the delayed effects of investments, we take a accumulation of investments over a number of years θ, with weights ξ, as our measure for investment: $M_{t_j} = \sum_{k=t-\theta}^{t} \xi_\theta I_{k_j}$.

For investment, there are data for nine years, 1979 to 1987. Investment series M have to be cumulated over a period of at least three or four years to get a reasonably smooth investment series. Cumulating over more than five years seems inappropriate, because a lot of this equipment is depreciated over a five year period, which indicates that banks count on replacement of this equipment after a five year period. It would not be suitable to assume an influence on productivity and capacity of investments that are likely to have been already scrapped. Probably the estimates are only moderately sensitive for additions of investments further back since the volume of this type of investments has grown quite considerably since the beginning of the eighties. Weighting investments from different years is problematic, because it should not only account for the large price drops in automation equipment, but also for price changes in application software, other office equipment and refurnishing, for the existence of learning effects over a number of periods and for scrapping. Reliable price indices for this type of investments are not available. There is thus no way a priori to determine an accurate scheme of weights. However, some experimenting with different schemes of weights, assuming a deflation per year of rates between 0% and 20% (on the assumption that the deflationary tendencies outweigh their inflationary counterparts), showed a very high rate of correlation between investment series weighed in different ways. Since the relative ordering of banks in the series seemed to be robust *vis a vis* these transformations, it was decided to take the unweighted sum of investments in automation equipment over five years as a measure for M.

Finally, some indicators for the technological level of a bank will be needed. Obvious candidates are the number of years a bank already uses a certain system in a specific years t. We shall use the number of years a bank uses back office automation equipment in 1988 and the number of years it uses front office automation equipment in 1988. In addition, an indicator has been constructed to express how long ago on average investments in automation equipment were done, the average investment lag, defined as $lag_T = \left(\sum_{t=1}^{7} tM_{T-t}\right)\Big/\left(\sum_{t=1}^{7} M_{T-t}\right)$. If two banks have invested the same amount, then this indicator will be higher for the bank that has invested first and is thus supposed to be on a more advanced technological level, due to learning effects.

Summing up, it seems most proper to use data series representing payment traffic as a measure for output. We use two series for output, mutations in current accounts Y_{1_j} and numbers of current accounts Y_{2_j}. Variable inputs are represented by three alternatives, total labour V_{0_j}, total labour input in payment traffic V_{1_j}, and total cashier's, counter and data processing labour input in payment traffic V_{2_j}. The last measure is probably more accurate, but the first might suffer less from errors of measurement. Investments are represented by a five year accumulation of series of investments in automation equipment M_{t_j}. For output and labour input, there are data for three years, 1985 to 1987, and for investment there are data for nine years, 1979 to 1987. All three our data series suffer from their particular deficiency. The proxy for capacity does not account for changes in utilization rate, the proxy for variable factor demand is total demand for one factor, instead of marginal demand for all factors, and the investment proxy is limited to investment in automation equipment and suffers from obscurities in the lag structure.

All together there are 58 banks for which there are data for all three years, 15 banks for which only data for 1985 and 1986 are available and 17 banks for which only data for 1986 and 1987 are in the data set. Thus there are all together 148 observations of changes of Y_{t_j} and V_{t_j} from one year to the next. There are two observations of large decreases in production output Y_{t_j}. These have been left out of the sample, because here it is likely that mechanisms of contraction of business are operative which are not captured by the model.

The size distribution of the banks is asymmetric. There are many small banks and about three very large banks in the sample (see Figure 7). Moreover, there is one bank that has invested disproportionately in automation equipment, even in relation to its size, and a bank that has not invested in automation between 1984 and 1986. These latter two banks have also been deleted from the sample.

The relationships between the variables chosen in this section are illustrated in a number of graphs in the appendix to this chapter. The first four graphs picture the variation in the data series for capacity and variable factor productivity; the next two illustrate the differences between various measures of capacity and variable inputs; the next six graphs represent the relationship between the key variables of the model, growth in production capacity and in variable factor productivity, and investment; the last six graphs illustrate the relationships between investment, on the one hand, and size and productivity, on the other hand.

4.6 Conclusion

The banking industry is one of the parts of the services sector that is most affected by the development of information technologies. The introduction of computer equipment in Dutch banks started already about three decades ago, but has gained momentum in the last ten years. Automation takes no longer place exclusively at bank's headquarters, but proliferates to local branch offices. The construction of networks for data exchange has facilitated and stimulated this trend toward decentralized automation. This trend has many consequences, not only for labour and productivity, but also for the product range, product quality, management and organization of the bank.

Part of this trend is reflected in our data set. The data cover some important steps in the automation of local branch offices of banks: automation of data processing, first at the back office and later at the front office. We presented figures on the introduction and diffusion of a number of systems, of which the most important were back office automation equipment, front office automation equipment and automatic teller machines. The diffusion of these systems seems to follow a well known pattern: an S-shaped diffusion curve. Beside data on technology adoption, we considered data on firm sizes and on costs. The distribution over firm sizes is heavily skewed to the left: there are many small firms and few large. Regression of innovation dates on size indicated that firm size in itself can only partly explain the date of adoption of back office automation or front office automation equipment. In particular, there are almost no large banks that adopt late, but there are quite some small banks that adopt early. Cost data showed that labour costs account for more that half of the total, and automation costs for less than 6% on average in 1987. Almost half of all costs are made in taking care of transfers of payments and administering payment accounts. Finally, the data were analysed to see whether there are economies of scale in banking. We found that on average there might be slight economies of scale. Looking at cost components, it turned out that there are economies of scale in a number of activities, like management, public relations, training, automation and mail. However, considering the costs of labour, especially of cashiers and counter clerks, there could be some diseconomies of scale.

The introduction of new technology is bound to have an influence on output and on both the quantity of employment and the quality of work. Figures do not adequately reflect changes in the quality of work or of production, nor do they say much about diversification of the product range. They might, however, contain some information on developments in labour productivity and output volume, in relation to investments in new technology, along the lines described by the models of chapter 3. From the data reviewed here, some series were selected to be used for testing these models. These series will serve as proxies for production capacity Y, for variable factor inputs V, for investments M, and for technological level τ. In the next chapter these data will be used to analyse the mechanics of technical change in banking. This is an hazardous venture. First of all, it is notoriously difficult in banking to calculate costs of products. Some services, like domestic payment transfers, are not even priced, but seen as a by-product of other operations. Under these circumstances one cannot expect bankers to calculate their odds as sharp as the industrialists in our models, when the marginal returns on investment are at stake.

Uncertainties about costs and benefits of investments in automation are substantial, which creates room for other arguments than financial return when decisions have to be taken. Rational decision making might take different forms in banking than in other sectors.

Secondly, banking in The Netherlands is dominated by a very small number of competitors that do not seem to engage in cut throat competition, although profits are not high to international standards. To approach the introduction of technology purely from the perspective of a firm struggling in a competitive market might be oversimplified. Maybe decisions concerning the speed of automation of banks are not taken solely at the level of the local bank, but also at a higher level. There might be a component of strategic behaviour at the level of the banking organization that does not show in the data.

Thirdly, technological change is a long term issue. Long term trends on the supply side are blurred by short term fluctuations on the demand side. Demand for banking products is volatile and largely independent of the pricing and the technology of the bank. The bank charges an interest margin to cover its costs, but the price of a banking product like a credit is largely determined by the cost of necessary raw materials: savings. These interest rates do not move under the control of the bank. Unfortunately, our time series are very brief and demand shifts might introduce some noise in the data.

Whatever the hazards may be, the attraction of the venture lies in the fact that the data set in itself has a large degree of detail, and in the notion that the introduction of information technologies gradually revolutionizes business in the banking industry, and the services sector at large, beyond recognition.

Appendix: some graphs illustrating the data to be used for model testing

Figures 11 and 12 show the output distribution of banks in 1986, measured in two ways, by number of mutations in current accounts Y_{1_j} and by total number of current accounts Y_{2_j} respectively. From both graphs, like from graph 6, it can be seen that the differences between banks are large: the largest bank is about ten times the size of the smallest bank. The graphs also show clearly that the size distribution is rather skewed: there are many small banks and a small number of relatively large banks.

Figures 13 and 14 show the distribution of variable factor productivity, first measured as number of mutations in current accounts Y_{1_j} per volume of processing, cashier and counter work V_{2_j}, and then measured as number of numbers of current accounts Y_{2_j} per volume of processing, cashier and counter work V_{2_j}. The graphs have the shape one might expect: a bell shaped distribution of productivity. The figures indicate that there are large differences in productivity between banks: the banks with the highest productivity are nearly 2.5 times as productive in producing mutations in accounts as the least productive banks. These large differences in productivity might be caused by differences in production technique, such that some banks use much more labour intensive techniques than others. Alternatively they might be caused by heterogeneity of output or labour inputs. Finally, they might be caused by errors of measurement, e.g. because banks register labour input into processing of payments in a different way.

Figures 15 and 16 show the relationships between the two proxies of output, Y_{1_j} and Y_{2_j}, and between different proxies of variable factors of production, on the one hand V_{1_j} and V_{2_j}, and on the other hand V_{0_j} and V_{2_j}. The growth rates of mutations are on average about 3 to 4% higher than the growth rates of the numbers of accounts. Growth rates of numbers and mutations of accounts show only a weakly positive relationship to each other. The measures of labour input V_{0_j} and V_{2_j} are only weakly correlated, but the measures V_{1_j} and V_{2_j} are strongly correlated.

Figures 17 and 18 thereafter plot growth in output versus growth in variable factor productivity, again measured in two ways. These are in fact the graphs that the model tries to explain. Output is measured as Y_{1_j} and Y_{2_j} respectively, and growth in variable factor productivity is measured as relative growth in output minus relative growth in variable inputs V_{2_j}. Clearly there is a large variation in combinations of expansion and productivity increase. Both graphs show many negative values for productivity growth. The fact that a large number of banks saw productivity drop is probably due to a slack in demand and thus underutilization of capacity. Such a fall in productivity is a feature that is not covered by the model of chapter 3.

Figures 19 to 22 illustrate the relationship between explanandum and explanans. The first two graphs show investment in automation equipment, cumulated over the years 1982 to 1986, and growth in output volume over 1986-1987, measured in two ways. The last two graphs show investment and growth in labour productivity over 1986-1987, where productivity equals output, here too measured in two ways, divided by variable factor productivity, using V_{2_j} as a proxy.

These graphs do not reveal strong partial correlations between the independent variable M and either of the dependent variables Y or Y/V.

Figures 23 to 26, relate investment between 1982 and 1986 to size in 1986, and investment between 1982 and 1986 to productivity in 1986. Investment precedes the measurement of size and productivity. There is a clear relationship between investment and size, but not between investment and productivity.

Figures 27 to 30, finally, picture the reversed relationship between investment and size and productivity: investment is measured in periods following the measurement of size and productivity. Investment is cumulated over the years 1985 to 1987, size is measured in 1985, by Y_{1_j} and Y_{2_j} respectively, and so is productivity in the year 1985, measured by Y_{1_j}/V_{2_j} and Y_{2_j}/V_{2_j}.

Changes in capacity and productivity and investment may be mutually dependent. Growth in capacity and productivity are caused by investment, but differences in capacity and productivity on their turn are likely to cause differences in investment. The relationship between size and subsequent investment is positive. Between productivity and subsequent investment there seems to be no relationship. Maybe the fact that investments could be cumulated over three years only is of influence here.

Figure 11

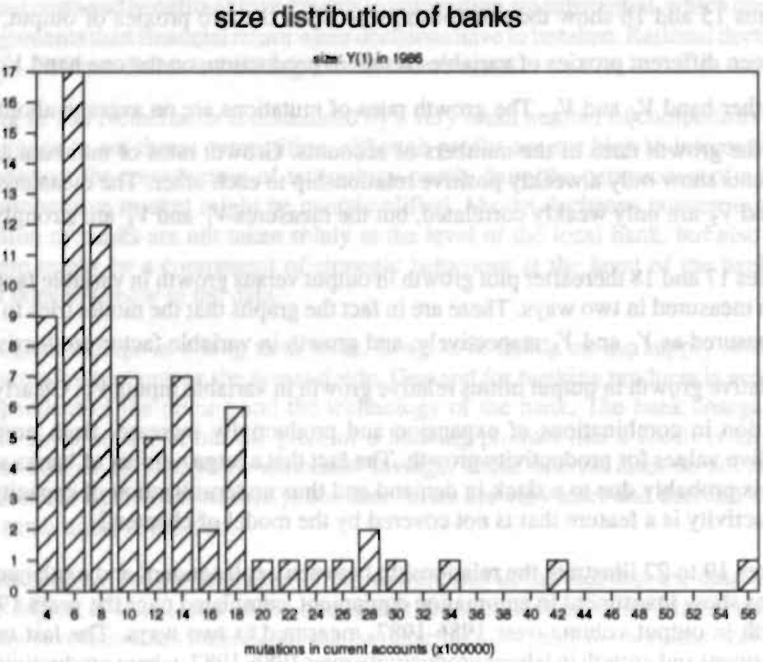


Figure 12

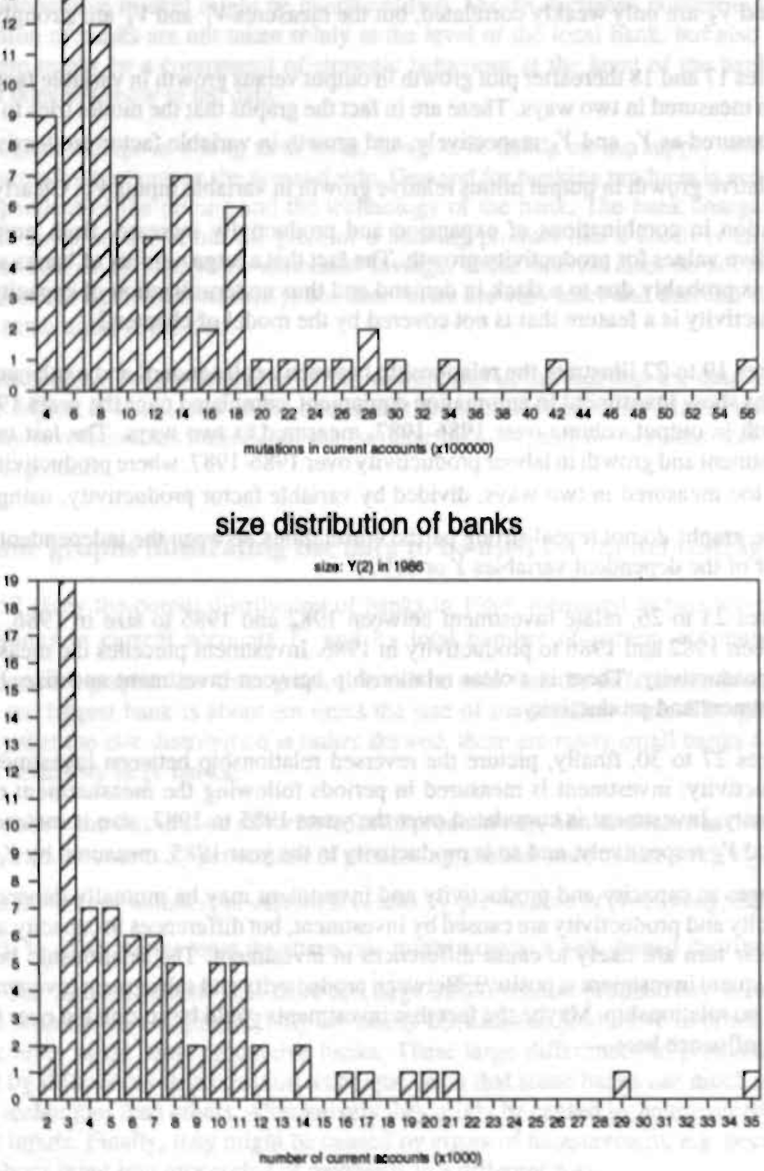


Figure 13

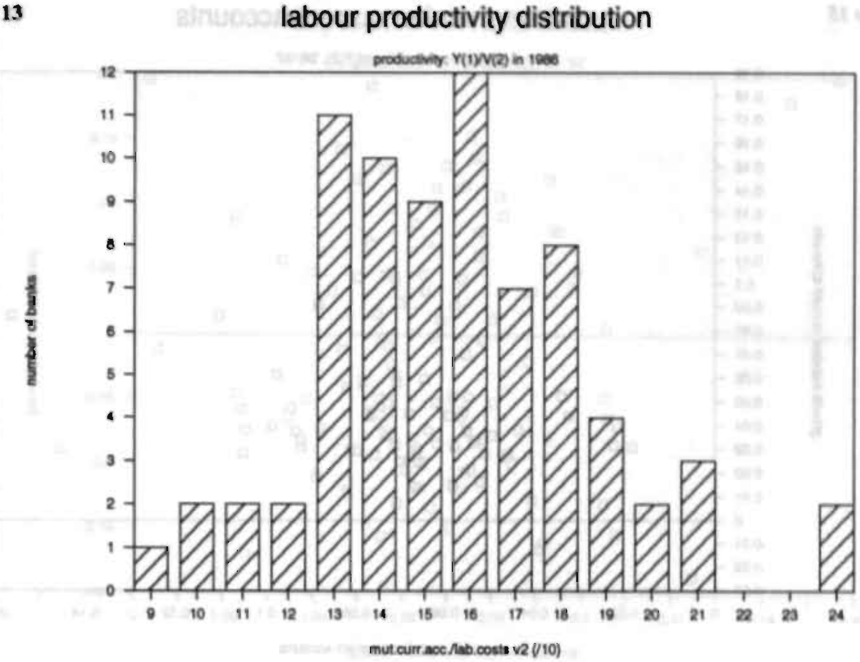


Figure 14

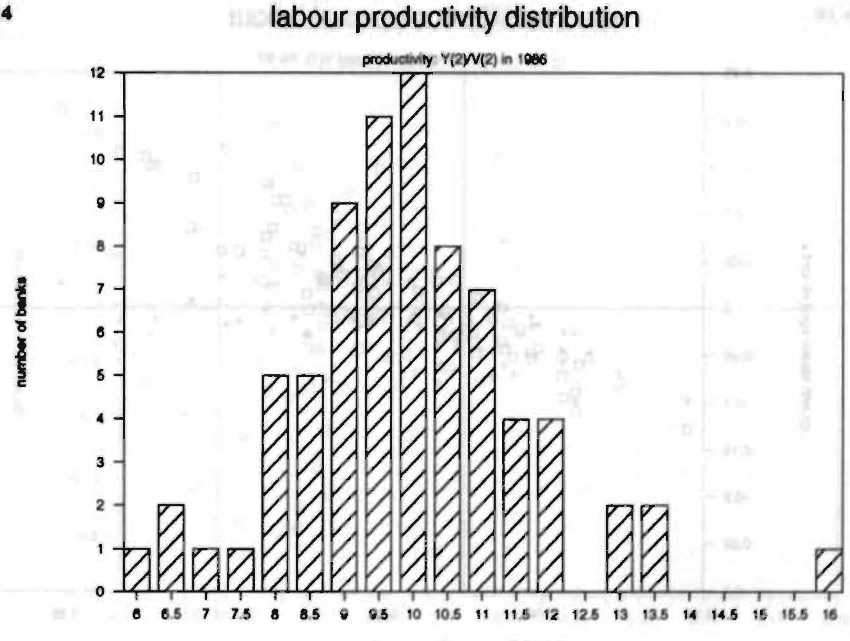


Figure 15

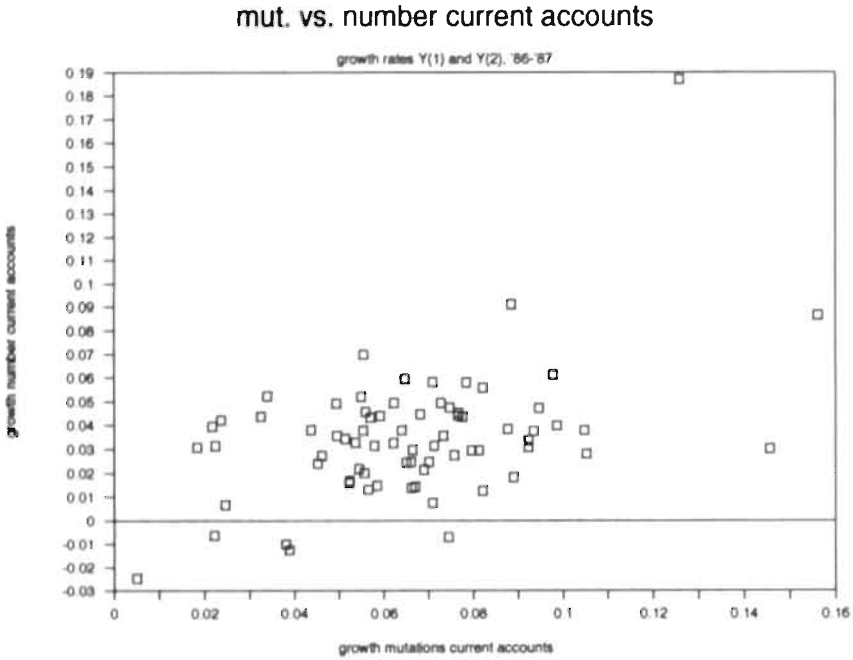


Figure 16

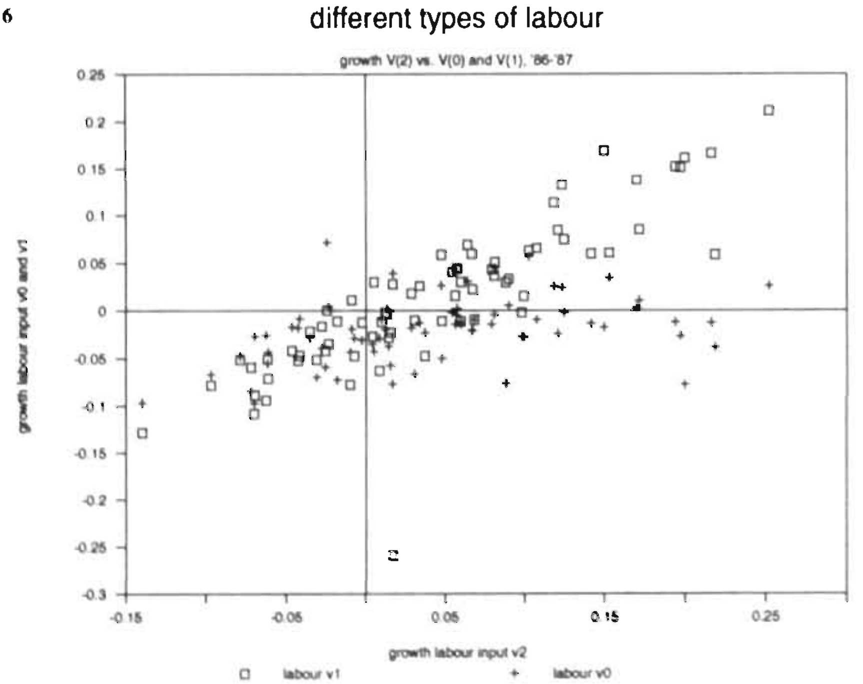


Figure 17

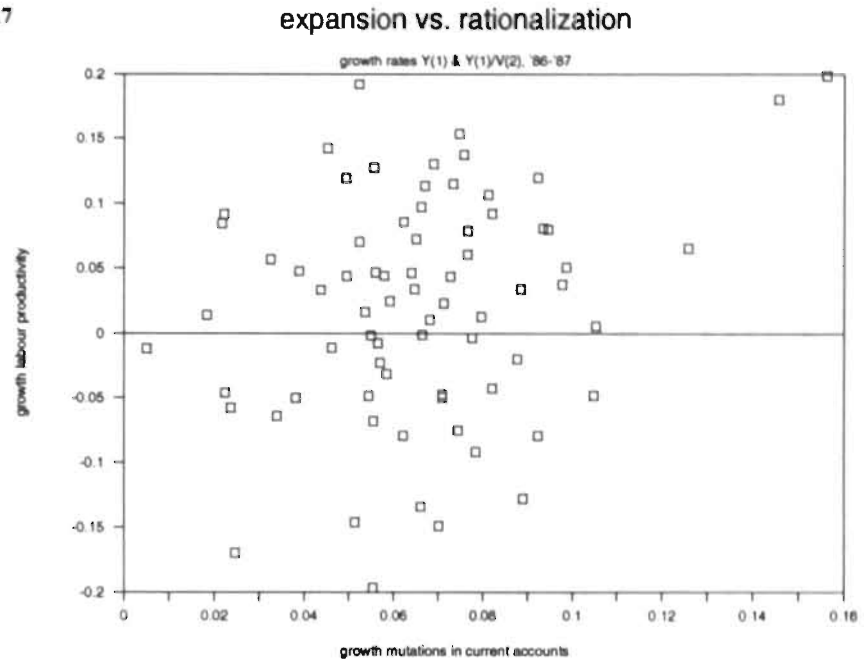


Figure 18

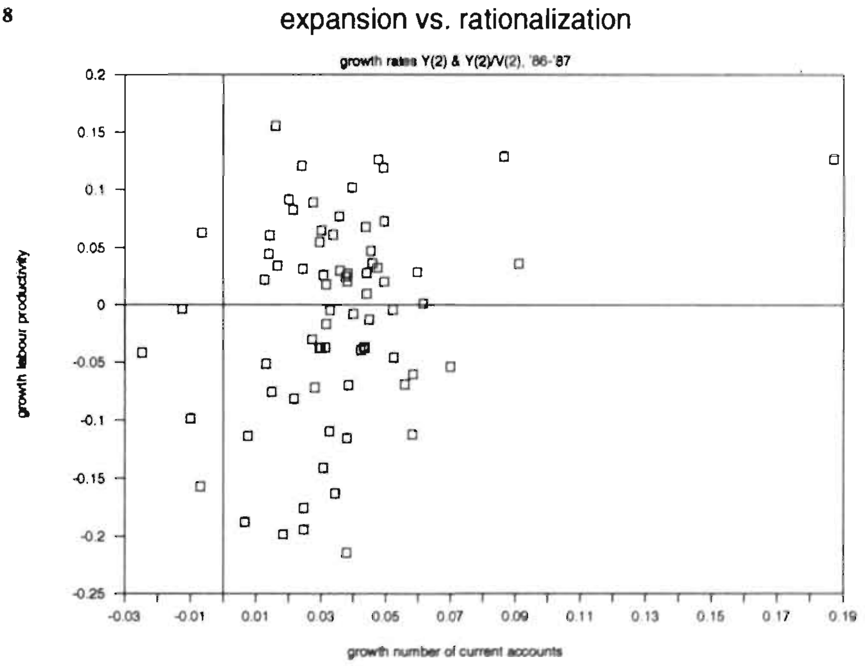


Figure 19

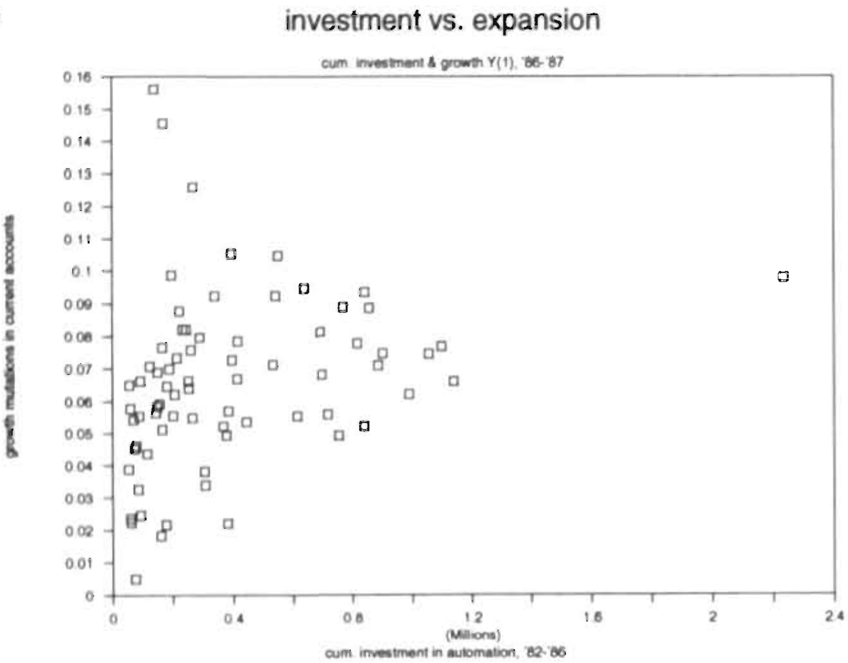


Figure 20

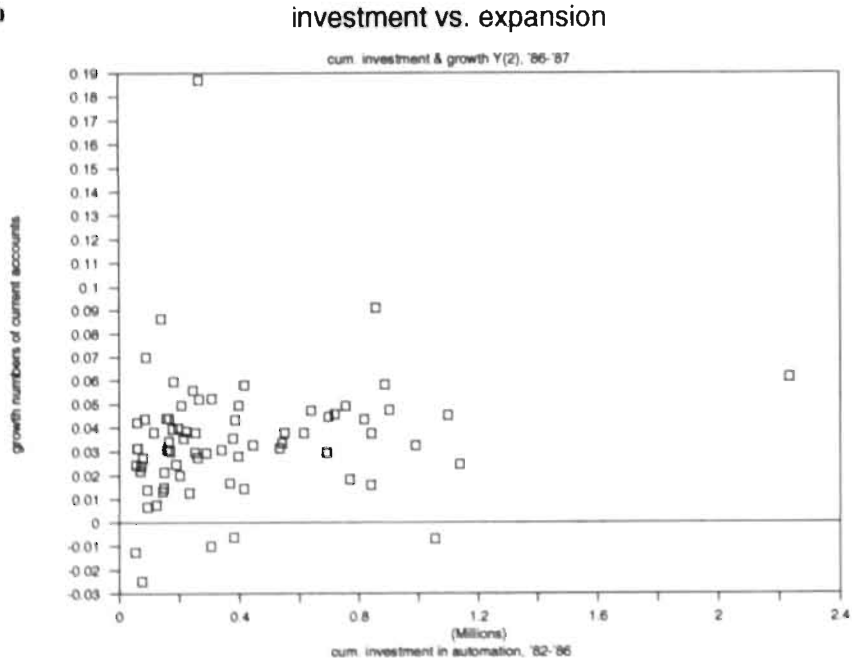


Figure 21

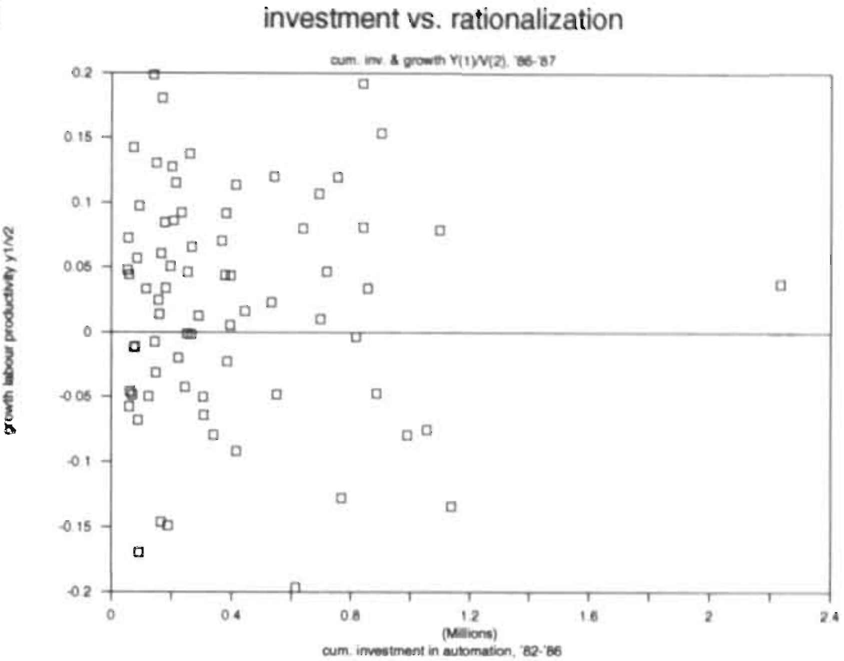


Figure 22

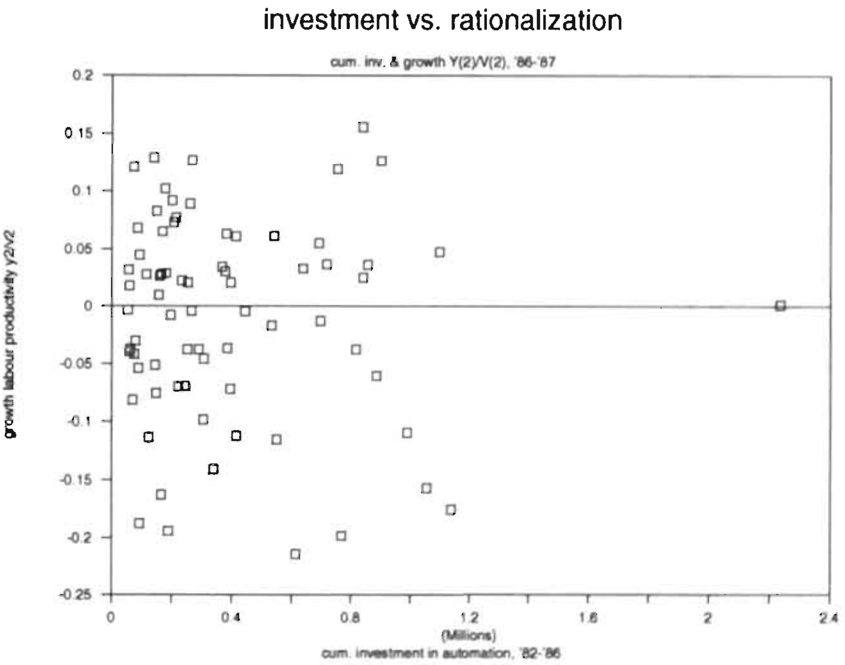


Figure 23

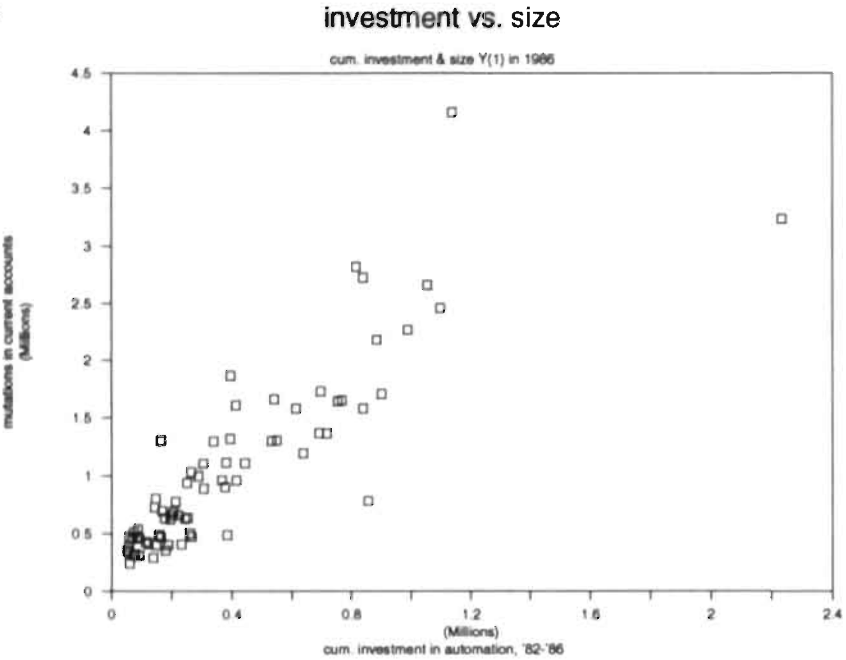


Figure 24

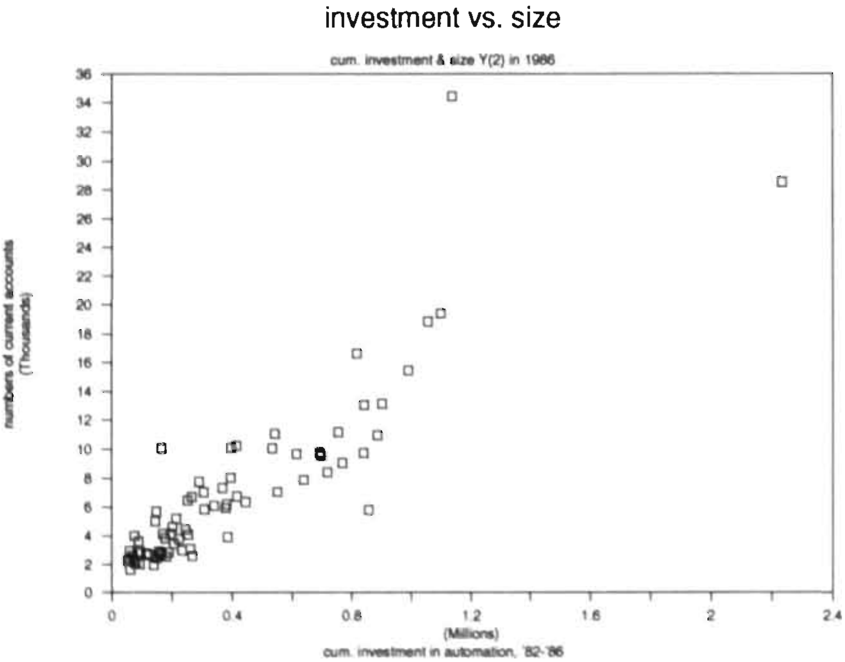


Figure 25

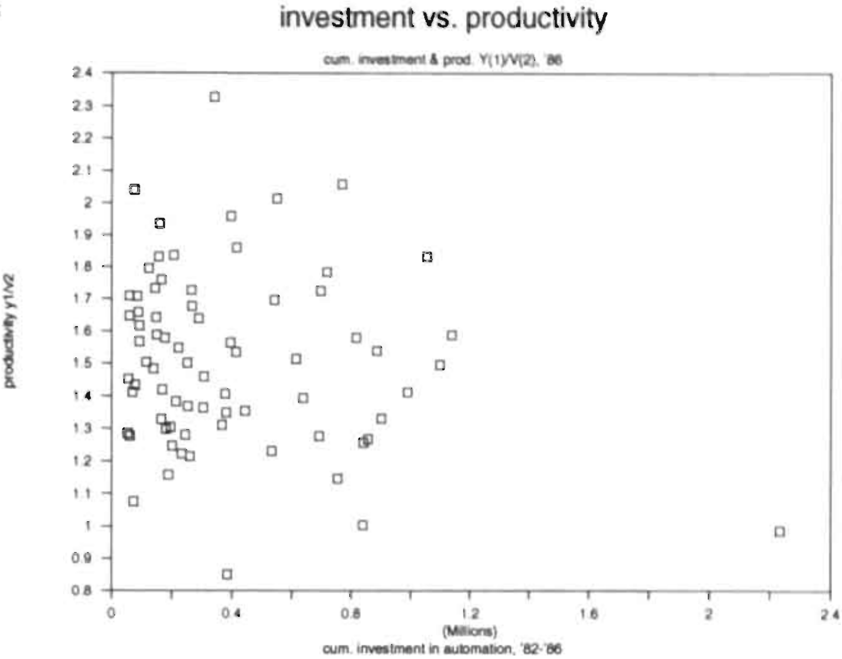


Figure 26

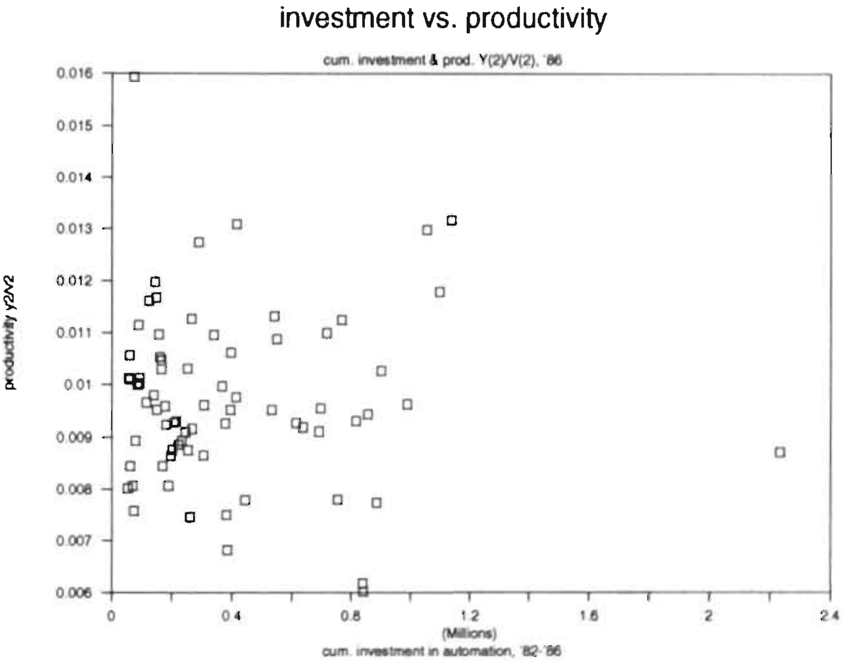


Figure 27

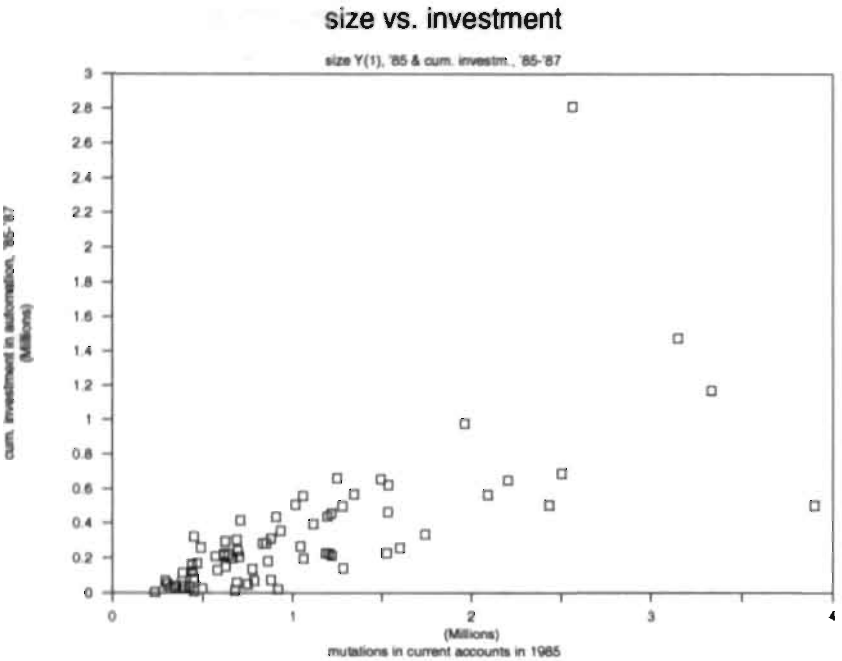


Figure 28

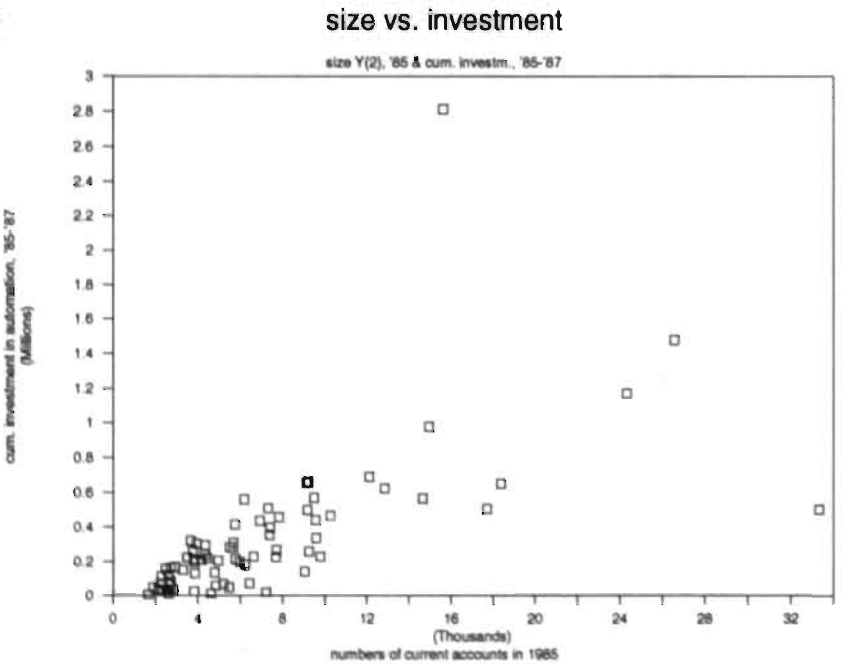


Figure 29

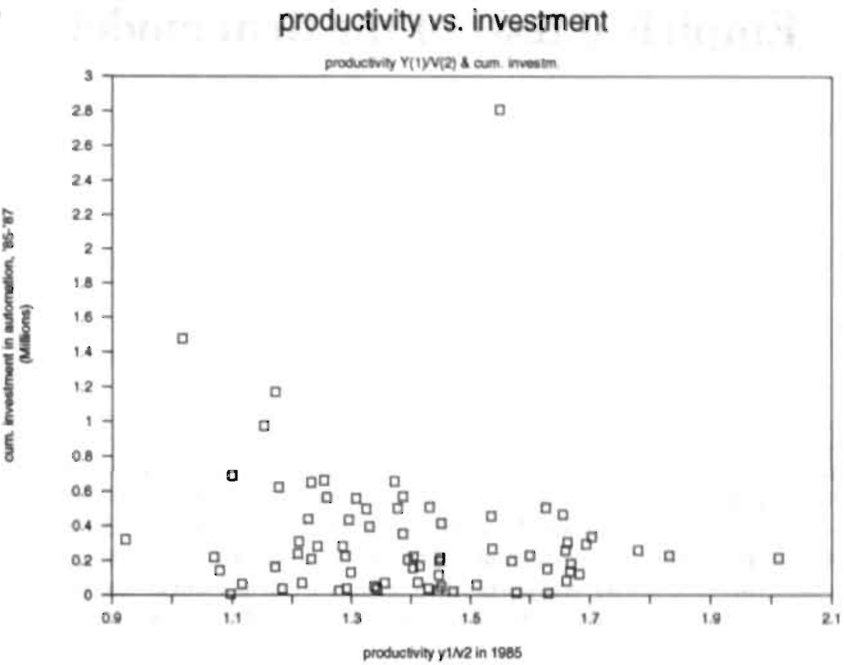
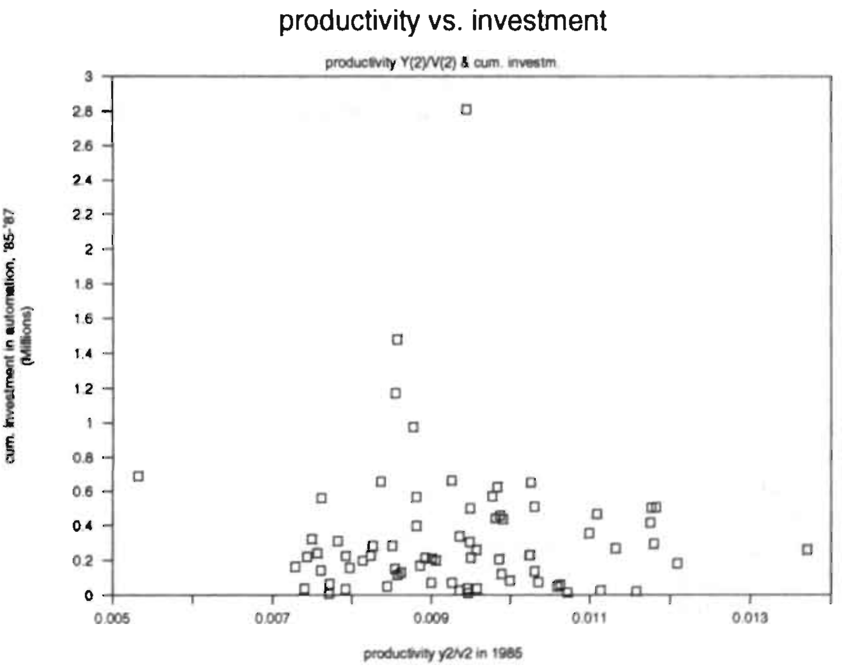


Figure 30



5. Empirical tests of the firm model

5.1 Introduction

In this chapter, the data set described in chapter 4 will be analysed with the help of the models of chapter 3. The models developed in chapter 3 describe the development of one single firm over a long period of time. The data set, however, comprises only short time series, but contains a lot of cross section information. This restricts the extent to which the model can be put to the test. Model predictions cannot be compared to actual long term developments. What can be done with cross section data, is to see to whether firms at one moment in time behave according to a pattern which can be predicted by a two-period version of the model. Firms proceed along their path of development, and one can try to discover whether there are common elements to these paths, and whether some firms are further along a common path than others.

The plan of the chapter is as follows. Before turning to the data, we take some time to reconsider and simplify the model. A number of hypotheses are developed that can be tested with the available short data series. This is done by abandoning the long term planning horizon and receding to the assumption of myopia. Using this simplified model, we analyse how inter firm differences in characteristics would work out in terms of inter firm differences in development of production capacity and productivity. Then we tackle the data and follow a strategy of testing ever less restrictive assumptions on the similarity of the development paths of firms. In a sense, we start out from testing the idea of a representative firm and proceed to allow for an increasing number of differences between firms, in order to find out to what extent all firms follow the same path.

5.2 A two-period model

In chapter 3, we considered the investment planning over time of a firm. The firm aims both at expansion and at rationalization. Two bench-mark states of the world were considered, one in which technology was unchanging, the other in which technical progress was pervasive, in the sense that every investment had a strong cumulative effect on productivity. To derive some hypotheses about the relationship between differences in firm characteristics and differences in behaviour, we reduce the general model to a two-period model, in which only one decision about a one time investment is taken. At the same time we introduce a more general specification of technical change, of which the two bench-marks mentioned above are special cases. The more general specification for the technical constraints allows not only for stationary technology and pervasive technical progress, but also for forms in between and for asymmetric technological progress, in which there is more technological opportunity to augment fixed factors than to save on variable factors, or *vise versa*. The starting point is once more the basic model of chapter 3, section 3. Depletion of technological opportunity and variation in prices are not considered. The model, reformulated for the myopic firm, is:

$$\text{Max}\, Z = PY_t - wV_t - rM_t \tag{1}$$

subject to:

$$\dot{Y}_t = Y_t^{\alpha_1} Y_0^{1-\alpha_1} \beta_t h(M_t) \qquad 0 \le \alpha_1 \le 1 \tag{2}$$

$$\dot{V}_t = V_t^{\alpha_2} V_0^{1-\alpha_2} (\beta_t - g(\beta_t)) h(M_t) \qquad 0 \le \alpha_2 \le 1 \tag{3}$$

In this chapter, a dot over a variable with index t indicates the change in this variable from period t to the next: $\dot{x}_t = x_{t+1} - x_t$. This convention keeps the notation here comparable to the notation in chapter 3, although we have moved from continuous to discrete time. The formulation of the technical constraints above encompasses both the bench-mark cases of chapter 3, sections 3 and 4. If $\alpha_1 = \alpha_2 = 0$, there is no technical change, and if $\alpha_1 = \alpha_2 = 1$, there is pervasive technical change. Intermediate values describe a mixture of technical change and plain expansion and substitution. We shall call the difference between α_1 and α_2 the bias in technological *opportunity*, not to be confused with the direction of or the bias in technical *change*, which depends on the firm's choice of $g(\beta)/\beta$, given price ratios, α_1 and α_2. If $\alpha_1 > \alpha_2$, then the current course of technological development gives more possibilities for technical progress in expansion of the production, than for decreasing variable factor requirements. Technological opportunity is thus biased towards augmentation of the capital stock. Conversely, if $\alpha_1 < \alpha_2$, technological opportunity is biased towards saving on the variable input flows. Finally, if $\alpha_1 = \alpha_2$, technological opportunity is neutral.

The characterization of technical change hinges on the presence of a cumulative effect, a learning effect. This cumulative effect cannot be discerned, if only two periods are considered. In a sense, however, the cumulative effect is being slipped into the model, by making the effect of investments on $\dot{Y}_t$ and $\dot{V}_t$, on growth of capacity and variable factor demand, a function of capacity Y_t and variable factor demand V_t, which are determined by firms' actions in the past.

It is assumed that there is no obsolescence. Therefore the objective function can be written as:

$$\text{Max}\, Z = PY_t + P\dot{Y}_t + \dot{P}Y_t + \dot{P}\dot{Y}_t - wV_t - w\dot{V}_t - \dot{w}V_t - \dot{w}\dot{V}_t - rM_t \tag{4}$$

Since prices are assumed constant, a number of the terms in the objective function drop out, because they are constant or zero. Substituting the constraints in the objective function, gives (suppressing time indices of β and M):

$$\text{Max}\, Z = PY_t^{\alpha_1} Y_0^{1-\alpha_1} \beta h(M) - wV_t^{\alpha_2} V_0^{1-\alpha_2} (\beta - g(\beta)) h(M) - rM \tag{5}$$

We write $c_1 \equiv Y_0^{1-\alpha_1}$ and $c_2 \equiv V_0^{1-\alpha_2}$, where c_1 and c_2 are both positive constants. Solving and writing $\gamma_t = Y_t/V_t$ for variable factor productivity yields:

$$\frac{\delta Z}{\delta \beta} = 0 \quad \Leftrightarrow \quad g'(\beta) = 1 - \frac{P Y_t^{\alpha_1} Y_0^{1-\alpha_1}}{w V_t^{\alpha_2} V_0^{\alpha_2}} = 1 - \frac{P}{w}\frac{c_1}{c_2} Y_t^{\alpha_1-\alpha_2} \gamma_t^{\alpha_2} \tag{6}$$

$$\frac{\delta Z}{\delta M} = 0 \quad \Leftrightarrow \quad h'(M) = \frac{r}{P Y_t^{\alpha_1} Y_0^{1-\alpha_1} \beta - w V_t^{\alpha_2} V_0^{1-\alpha_2} (\beta - g(\beta))} = \frac{r}{P c_1 Y_t^{\alpha_1} \beta - w c_2 Y_t^{\alpha_2} \gamma_t^{-\alpha_2} (\beta - g(\beta))} \tag{7}$$

Equation (6) indicates that the optimal direction of the investment depends on the bias in technological opportunity, the difference between α_1 and α_2. A higher α_1 and a lower α_2 boasts β and curbs investment towards expansion. Equation (7) says that optimal total investment in progress grows with technological opportunities for expansion α_1. It depends on the sign of $\beta - g(\beta)$, whether larger opportunities in variable input saving α_2 lead to a higher budget: if $g(\beta)$ exceeds β, investments rise when α_2 rises. Like in chapter 3 (see equations (1) and (2) there), we impose some restrictions on the trade-off function $g(\beta)$ and on the investment function $h(M)$. The trade-off function is downward sloping ($g'(\beta) < 0$), and the investment function is upward sloping ($h'(M) > 0$).

By analyzing equations (6) and (7), differences in firm behaviour can be traced, as a function of technical possibilities and differences in firm characteristics, notably in size Y_t and variable factor productivity γ_t. Taking the derivatives of equations (6) and (7) with respect to Y_t and γ_t, we consider the differences in slope of $g(\beta)$ and $h(M)$ in the optimum, for different firm sizes and variable factor productivities. First consider the derivative of $g'(\beta)$ with respect to production:

$$\frac{dg'(\beta)}{dY_t} = -\frac{P}{w}\frac{c_1}{c_2} \gamma_t^{\alpha_2} (\alpha_1 - \alpha_2) Y_t^{\alpha_1-\alpha_2-1} \tag{8}$$

If technological opportunity is biased, then a larger scale of production causes investments to be more curbed in the direction of the bias:

$$\alpha_1 > \alpha_2 \quad \Leftrightarrow \quad \frac{dg'(\beta)}{dY_t} < 0 \quad \Leftrightarrow \quad \frac{d\beta}{dY_t} > 0 \tag{9}$$

The optimal direction of investment changes as scale of production increases. The direction of the change, however, depends on the bias in technological opportunities: if there is more opportunity for 'variable factor saving technical change' than for 'capital stock augmenting technical change' ($\alpha_1 < \alpha_2$), then the larger the firm, the larger the share of investment that will be directed towards variable input saving, given equal variable factor productivity γ_t. If the bias is in the opposite direction ($\alpha_1 > \alpha_2$), then the larger the firm, the more investment will be directed towards fixed factor augmentation. In general, the *larger* the firm, the *more biased* technical change. The result holds, because larger firms can reap more profits from the cumulative effects of technical change. They are more able to take advantage of the cumulative effect of investment.

Secondly, take the derivative of $g'(\beta)$ with respect to variable factor productivity. The higher a firm's variable factor productivity, the more it will direct investment towards expansion. More efficient firms grow faster.

$$\frac{dg'(\beta)}{d\gamma_t} = -\frac{P}{w}\frac{c_1}{c_2} Y_t^{\alpha_1-\alpha_2} \alpha_2 \gamma_t^{\alpha_2-1} < 0 \quad \Rightarrow \quad \frac{d\beta}{d\gamma_t} > 0 \tag{10}$$

Inequality (10) holds for all permitted values of α_1 and α_2.

Thirdly, consider the derivative of $h'(M)$ with respect to production capacity:

$$\frac{dh'(M)}{dY_t} = -r\left(\frac{Pc_1\alpha_1 Y_t^{\alpha_1-1}\beta - wc_2\alpha_2 Y_t^{\alpha_2-1}\gamma_t^{-\alpha_2}(\beta - g(\beta))}{\left(Pc_1 Y_t^{\alpha_1}\beta - wc_2 Y_t^{\alpha_2}\gamma_t^{-\alpha_2}(\beta - g(\beta))\right)^2}\right) \tag{11}$$

A larger firm will spend a higher investment budget under the following condition:

$$\frac{dM}{dY_t} > 0 \quad \Leftrightarrow \quad \frac{dh'(M)}{dY_t} < 0 \quad \Leftrightarrow \quad \alpha_1 Pc_1 Y_t^{\alpha_1}\beta - \alpha_2 wc_2 Y_t^{\alpha_2}\gamma_t^{-\alpha_2}(\beta - g(\beta)) > 0 \quad \Leftrightarrow \tag{12}$$

$$Y_t^{\alpha_1-\alpha_2} > \left(\frac{\alpha_2}{\alpha_1}\right)\left(\frac{w}{P}\frac{c_2}{c_1}\right)\left(1 - \frac{g(\beta)}{\beta}\right)\gamma_t^{-\alpha_2}$$

This condition is fulfilled, if it is optimal to direct investment more towards variable input flow saving, $g(\beta) > \beta$, because in that case the left hand side is positive and the right hand side is negative. The condition can also be fulfilled, if $g(\beta) < \beta$, but then it is less straightforward to determine (and interpret) the necessary conditions, because the model is not invariant to changes in the split of value terms into a volume and a price component.

Finally, take the derivative of $h'(M)$ to variable factor productivity:

$$\frac{dh'(M)}{d\gamma_t} = -r\left(\frac{wc_2 Y_t^{\alpha_2}\alpha_2\gamma_t^{-\alpha_2-1}(\beta - g(\beta))}{\left(Pc_1 Y_t^{\alpha_1}\beta - wc_2 Y_t^{\alpha_2}\gamma_t^{-\alpha_2}(\beta - g(\beta))\right)^2}\right) \tag{13}$$

The condition for a larger investment budget as productivity grows is:

$$\frac{dM}{d\gamma_t} > 0 \quad \Leftrightarrow \quad \frac{dh'(M)}{d\gamma_t} < 0 \quad \Leftrightarrow \quad \beta > g(\beta) \tag{14}$$

If it is optimal to spend resources relatively more on expansion than on rationalization, then the higher a firm's variable factor productivity, the more it will invest. If, by contrast, it is optimal to spend resources more on rationalization, then the higher a firm's variable factor productivity, the less it will invest. The conclusions are summarized in Table 1.

Table 1: The relationships between changes in size and productivity, and amount and direction of investment.

	$\Delta\beta$	ΔM
ΔY	$\uparrow$ *if* $\alpha_1 > \alpha_2$ $\downarrow$ *if* $\alpha_1 < \alpha_2$	$\uparrow$ *if* $Y^{\alpha_1-\alpha_2} > \left(\frac{\alpha_2}{\alpha_1}\right)\left(\frac{w}{P}\frac{c_2}{c_1}\right)\left(1-\frac{g(\beta)}{\beta}\right)\gamma^{-\alpha_2}$ $\downarrow$ *if* $Y^{\alpha_1-\alpha_2} < \left(\frac{\alpha_2}{\alpha_1}\right)\left(\frac{w}{P}\frac{c_2}{c_1}\right)\left(1-\frac{g(\beta)}{\beta}\right)\gamma^{-\alpha_2}$
$\Delta\gamma$	$\uparrow$	$\uparrow$ *if* $\beta > g(\beta)$ $\downarrow$ *if* $\beta < g(\beta)$

Table 1 shows that the direction of technical progress changes monotonously with size and with productivity. The investment budget, however, does not necessarily display a monotonous pattern, because the sign of the derivatives of $h'(M)$ depends on size and productivity, as well as on the direction of technical progress.

First consider the bottom row of Table 1. If productivity goes up, *firms direct their investments* more toward expansion. Moreover, as firms are more productive they invest more, provided that the optimal direction of investment is toward expansion. By contrast, as firms are more productive they invest less, if the optimal direction of investment is toward rationalization. If firms differ in productivity only, and if low productivity firms invest in rationalization ($\beta < g(\beta)$) and high productivity firms invest in expansion ($\beta > g(\beta)$), the relationship between productivity and amount of investment may be downward sloping up to the value of productivity for which in the optimum $\beta = g(\beta)$, and upward sloping after.

Now consider the top row of Table 1. Larger firms direct their investments more in the direction of the bias of technological opportunities. It is easy to see that for large firms of a certain productivity the amount of investment varies positively with size. If $\alpha_1 > \alpha_2$, there is a specific size above which the upper condition holds. If $\alpha_1 < \alpha_2$, there is a specific size above which $\beta < g(\beta)$, and thus the lower condition is violated. The relationship between size and investment at lower levels of size is indeterminate and may be not monotonous.

Leaving the theoretical possibility of a locally negative relationship between investment and size for what it is, but restricting ourselves to the more usual pattern, we can summarize our findings in four hypotheses:

1. Large firms invest in a more biased direction than small firms.
2. Large firms generally spend more on investment.
3. Firms with more output per unit of variable input invest more in expansion.
4. Firms with more output per unit of variable input invest more, if it is optimal to invest relatively more in expansion and less in rationalization.

5.3 Hypotheses and tests

The models of chapter 3 and of the previous section describe how rational firms act in an environment that is characterized by a possibility for a trade off between expanding production capacity and increasing variable factor productivity. The firm is assumed to choose an optimal mix of expansion and rationalization of production, taking factor prices into account, and to choose an optimal level of progress, depending on the marginal productivity of investment. Assume there are different firms j to which the model applies. The general specification of the technical constraints that firm j faces at time t is:

$$\dot{Y}_{t_j} = Y_{t_j}^{\alpha_1} Y_0^{1-\alpha_1} \beta_{t_j} h_j(M_{t_j}) \tag{15}$$

$$\dot{V}_{t_j} = V_{t_j}^{\alpha_2} V_0^{1-\alpha_2} (\beta_{t_j} - g_j(\beta_{t_j})) h_j(M_{t_j}) \tag{16}$$

The technical constraints of firm j are determined by two functions, $g_j(\beta)$ and $h_j(M)$, and the parameters α_1 and α_2. Every firm j makes a choice for the value of its instruments β_j and M_j. The choices β_j and M_j of different firms j result in capacity growth and variable factor increase. If α_1 and α_2 were equal to one, then these would be reflected in Figures (17) and (18) of chapter 4 above. With respect to the function $h_j(M_j)$ and the optimal choice M_j, one out of four possibilities has to apply:

1. Every single firm faces the same technical constraint function: $h_j(M) = h(M)$. At the same time, every firm is faced with the same choice problem, the same parameters, prices and possibilities. Therefore all firms take the same decision. The variation in the outcomes which is apparent from Figures (17) and (18) in the last chapter is due to some process which is independent of the determination of M_j, such that $M_j = M + \varepsilon_j$, where the ε_j is a random disturbance. In view of the outcomes of section 2 it is unlikely that the banks in our data set have all made the same choices. It was shown above that firms of different size or variable factor productivity will choose different values M_j.

2. Every firm faces the same constraint function: $h_j(M) = h(M)$, but despite of that firms make different choices M_j. Figures (17) and (18) are not the result of one optimal choice, with some random disturbances, but results from a range of optimal choices. Underlying this choice process, however, there is one stable technical constraint.

   Firms can be expected to choose a different M_j, although the constraint function $h(M)$ is the same for all firms, if there are differences in current output capacity or productivity, as has been shown in section 2. Firms will also choose differently if they face different prices or expect different price developments.

   If the function $h(M)$ looks the same for all firms, but different firms choose different values M_j, then, on the one hand, it must be possible to find the outlines of the function $h(M)$, on which every firm chooses one point. On the other hand, it might be possible to detect which differences in firm characteristics account for observed differences in choice M_j.

3. Firms each face a different function $h_j(M)$, and therefore make different choices M_j. Every firm may act as described by the model, and for each firm the constraint function $h_j(M)$ may be stable over time. The shape of this function, however, can differ across firms. Thus firms make different choices, despite of the fact that they all choose rationally.
   One important reason why the shape of $h_j(M)$ would differ over firms was advanced in section 5 of chapter 3: $h_j(M)$ could be dependent on the present technological level of the firm, on technical abilities and on perceived opportunities, represented there by the variable τ.
   Because only cross section data are available, we cannot test for stability of any function $h_j(M)$ for firm j over time, but only for similarity across firms. If the constraint function would differ across firms, then it might be possible to relate this variation to variety in firm characteristics which mirror technological level, e.g. by specifying $h_j(M) = h(M, \tau_j)$.

4. A last possibility is that the model does not describe the choice process of the firm to any reasonable extent. Investment is not related to expansion and rationalization as subsumed in the model. Other mechanisms dominate the course of developments.

Similarly, with respect to the function $g_j(\beta_j)$ and the optimal choice β_j, one out of four possibilities has to apply.

1. Every single firm faces the same technical constraint function: $g_j(\beta) = g(\beta)$. Also the other parameters of the choice problem are the same, and thus all firms take the same decision: $\beta_j = \beta + \varepsilon_j$. Again, it is unlikely that all firms make the same choice, if they differ in size or in variable factor productivity. All firms would invest in the same direction, only if $g(\beta)$ were non-differentiable (have a corner, where both $\beta, g(\beta) > 0$).

2. Every firm faces the same constraint function, $g_j(\beta) = g(\beta)$, but firms make different choices β_j. Referring again to section 2, it was shown that differences in current output capacity or productivity can lead to different choices β_j. If the function $g(\beta)$ is the same for all firms, but different firms choose different values β_j, we might find the outlines of this function $g(\beta)$. Also, it might be possible to detect what differences in firm characteristics account for observed differences in choice β_j.

3. Firms each face a different function $g_j(\beta)$, and therefore make different choices β_j. Also here differences in level of technological development might be of influence.

4. Finally, it could be that the model does not describe the choice process of the firm accurately enough and that other mechanisms dominate the course of developments.

There is one more dimension that could be considered: it could be assumed that all firms face the same parameters α_1 and α_2, or that the α's would be different for all firms. Because the α's mainly affect the amount of technical progress which results from an investment M by a firm, it is natural to consider them together with the function $h(M)$. Summarizing the above, one can distinguish between two functions that determine the technical constraints of any firm, and to each of these, one out of four descriptions must apply. Together this makes 16 possible states of the world that could provide the background to Figures (17) and (18). These possibilities can be summarized in a matrix:

Table 2: Possible extent of similarity of different firms.

		same $h(M)$, α_1 and α_2		diff. $h(M)$, α_1, α_2	no $h(M)$
		same choice M	diff. choice M	diff. choice M	no M
same constr. $g(\beta)$	same choice β	1a	1b	1c	1d
	diff. choice β	2a	2b	2c	2d
diff. constr. $g(\beta)$	diff. choice β	3a	3b	3c	3d
no $g(\beta)$	no β	4a	4b	4c	4d

It follows from our theoretical analysis (see Table 1 above) that some states are more likely to occur than others. Given different sizes and productivities of firms, we expect different choices for the instruments β and M. Given different levels of technological development we expect the shapes of the functions $g(\beta)$ and $h(M)$ to be firm dependent. However, we do not know a priori whether the differences between firms will be substantial, and whether these differences in firm characteristics can be related explicitly to differences in constraints and in choices.

What *cannot* be tested with our data base, is the validity of the model as a description of the development of a single firm over time, and thereby the accuracy of the assumptions about the operative decision mechanisms. For this, time series are required. What we *can* test, given these data, is similarity of choices and of constraints across firms. It is worthwhile to test for this similarity of constraints and decisions over firms, because it adds to the usefulness of the model, maybe even to its value as a tool for forecasting, if it appears that constraints and choices show some robustness across firms.

In the remainder of this section, we shall move systematically through Table 2, analysing the various combinations of hypotheses. We start on the left side of the matrix (column 1) with the most restrictive hypothesis concerning the amount of investment, and on the top (row a) with the most restrictive hypothesis concerning the direction of progress. As any of these hypotheses has to be rejected, we move either down or to the right. Thus we evaluate the hypothesis of a constant amount of investment M, similar for all banks, against the hypothesis of a varying amount of investment M_j. If the hypothesis of a fixed amount M is not supported by the evidence, the first column in the matrix can be discarded and we move to column 2. Then we consider the hypothesis of a choice of direction β, similar for each bank, which is tested against the hypothesis of a varying β_j. If this hypothesis is not in line with the evidence, the first row can be skipped and we go to row b. In this way one can continue in the direction of the lower right hand corner.

5.3.1 Same constraints, same decisions

The most restrictive hypothesis on the amount of investment is represented in the first column: the choice for the amount of investment M is the same for each and every bank, except for an error term. This implies that there is a common function $h(M)$, and that α_1 and α_2 do not vary over firms. Furthermore it is assumed that the variation in M that we see from the data is a random disturbance which could be due to measurement errors and other factors not systematically represented in the data.

On the basis of our analysis, we expect the choice of investment to vary with size: large firms, given a common variable factor productivity, should spend more on investment, because they reap more benefits from investment due to scale economies. We also expect investment to vary with variable factor productivity: firms with a higher variable factor productivity, given a certain size, should invest more if expansion is more profitable than rationalization, $\beta > g(\beta)$, and vice versa when rationalization is more profitable, $\beta < g(\beta)$. Because it was also shown that a higher variable factor productivity increases the profitability of expansion β, the first case is more likely to prevail for high productivity firms than the second case. If systematic variation in investment can be demonstrated, we are able to reject the cells in the first column and move to the right.

To analyse the relationships between size and investment, we need investment figures from a date not earlier than the period from which we have size and productivity figures. Our earliest *data on output and labour demand are from 1985. Thus we consider investments cumulated over* 1985, 1986 and 1987, data for Y_{1_j} and Y_{2_j} from 1985, and data for V_{0_j}, V_{1_j} and V_{2_j} also from 1985. We run the following regression:

$$\sum_{t=1985}^{1987} M_{j_t} = b_0 + b_1 Y_{i_j} + b_2 \frac{Y_{i_j}}{V_{k_j}} \qquad i = 1,2 \quad k = 0,1,2 \tag{17}$$

The results of this regression are listed in Table 3 below. It was clear from Figures (27) and (28) in chapter 4 that size is a fairly accurate predictor of investment. This is confirmed by the outcomes of the regressions: size is highly significant. Contrary to expectations, however, most of the regressions indicate that investment does not vary systematically with our measures of productivity, except for the fifth one reported. If there is any systematic relationship between investment and productivity, the regression results indicate that it is more likely to be negative than positive: a higher productivity leads to less investment in automation equipment. From Table 1, section 2, one can see that this could occur in our model, if β would on average be about equal to, or only slightly smaller than, $g(\beta)$.

Since investment shows systematic variation, we drop the hypothesis of the first column in the matrix of Table 2, and move to the right. Now consider the first row. The hypotheses represented here say that there is a common function $g(\beta)$, and a common choice of direction β. Thus the variation in direction of progress would be random. To analyse these hypotheses on the relationship between the growth of productivity and the growth of variable factor productivity, we cannot go about in the same way as above. The problem is that the direction of technical change cannot be measured independently, without knowing the values of α_1 and α_2.

Table 3: Results of estimation of investment equation (17).

output	input	b_0	b_1	b_2	R^2
Y_1	V_0	-50556	.371	-10.167	.468
		(-.241)	(8.006)	(-.023)	
	V_1	269710	.368	-330780	.484
		(1.187)	(8.142)	(-1.478)	
	V_2	143880	.368	-139800	.474
		(.602)	(8.018)	(-.859)	
Y_2	V_0	165900	45.015	-52768	.407
		(.891)	(7.176)	(-.863)	
	V_1	460630	47.563	-71992000	.441
		(2.208)	(7.648)	(-2.227)	
	V_2	376790	46.008	-40077000	.422
		(1.582)	(7.371)	(-1.573)	

Number of observations: 73; t-statistics in parentheses.

The theoretical analysis in section 2 suggests that the direction of technical progress $g(\beta)/\beta$ should vary with size and productivity, under the condition that $g(\beta)$ is concave and downward sloping. There is no guarantee that this condition is fulfilled, or that there would not be other constraints operative. A number of hypotheses concerning the direction of technical progress could be introduced and translated into extra constraints. Consider the following four which are selected because they have a straightforward economic interpretation:

1. The function $g(\beta)$ is concave and downward sloping, and can be locally approximated by a second degree polynomial: $g(\beta) = a_0 + a_1\beta + a_2\beta^2$. This specification is chosen for mathematical convenience. The main restriction is that $a_2 < 0$. The function $g(\beta)$ should be concave for all relevant values of β. For all positive β, $g(\beta)$ is certainly concave, if $a_1 < 0$, but if $a_1 > 0$, the function is also concave for all β larger than some critical value. If a_1 is estimated to be positive, it should be checked whether the estimated β's are in the concave part of the function. In equilibrium, $g'(\beta) = 1 - (P/w)(c_1/c_2)(Y^{\alpha_1}/V^{\alpha_2})$, as is shown in section 2.

2. The function $g(\beta)$ is non-differentiable in one point: it is kinked. This would mean that technical progress has a fixed direction, independent of price movements, size and productivity: $g(\beta)/\beta = c$.

3. The rate of expansion is constrained for some exogenous reason (say as a strategic objective): $\dot{Y}/Y = c$.

4. The rate of productivity increase is constrained or determined exogenously: $\dot{Y}/Y - \dot{V}/V = c$ (e.g. because the wage rate increases exogenously).

The function $h(M)$ should be upward sloping and concave. To pursue the analysis any further, it cannot be avoided to introduce a specification for $h(M)$. Assume that $h(M) = M^{\delta_1}$, where $\delta_1 < 1$ (a scale parameter is already included in the expressions specifying the direction of progress). Given these four alternative assumptions on the direction of technical change, and the assumption on the specification of $h(M)$, we end up with four possible models of firm behaviour. The objective function of the firm is always the same (dropping time indices):

$$\max_{\beta, M} Z = P\dot{Y} - w\dot{V} - rM \tag{18}$$

The constraints that were assumed so far were:

$$\dot{Y} = Y^{\alpha_1} c_1 \beta h(M) \tag{19}$$

$$\dot{V} = V^{\alpha_2} c_2 (\beta - g(\beta)) h(M) \tag{20}$$

The first extra constraint is:

$$h(M) = M^{\delta_1} \tag{21}$$

The four possible second extra constraints are:

1. $$g(\beta) = a_0 + a_1 \beta + a_2 \beta^2 \tag{22}$$

2. $$1 - \frac{\dot{V}/V^{\alpha_2}}{\dot{Y}/Y^{\alpha_1}} = c \tag{23}$$

3. $$\frac{\dot{Y}}{Y} = c \tag{24}$$

4. $$\frac{\dot{Y}}{Y} - \frac{\dot{V}}{V} = c \tag{25}$$

These constraints can be substituted into the original problem. The results in terms of $\dot{Y}$ and $\dot{V}$ are listed below (they can be easily derived). Maximization of the objective function, given these four constraint sets, gives four expressions for equilibrium investment M^* which are also listed. Here x_1 and x_2 are parameters, different ones in every case.

For the *first* set of constraints:

$$\dot{Y} = Y^{\alpha_1} \frac{1}{2a_2} \left(1 - a_1 - \frac{P}{w} \frac{c_1}{c_2} \frac{Y^{\alpha_1}}{V^{\alpha_2}} \right) M^{\delta_1} \tag{26}$$

$$\dot{V} = V^{\alpha_2}\left(-a_0 + \frac{(1-a_1)^2}{4a_2} - \frac{1}{4a_2}\left(\frac{P}{w}\frac{c_1}{c_2}\frac{Y^{\alpha_1}}{V^{\alpha_2}}\right)^2\right)M^{\delta_1} \tag{27}$$

$$M^{*} = \left\{\frac{\delta_1}{r}\left(P c_1 Y^{\alpha_1}\frac{1}{2a_2}\left(1 - a_1 - \frac{P}{w}\frac{c_1}{c_2}\frac{Y^{\alpha_1}}{V^{\alpha_2}}\right) - w c_2 V^{\alpha_2}\left(-a_0 + \frac{(1-a_1)^2}{4a_2} - \frac{1}{4a_2}\left(\frac{P}{w}\frac{c_1}{c_2}\frac{Y^{\alpha_1}}{V^{\alpha_2}}\right)^2\right)\right)\right\}^{\left(\frac{1}{1-\delta_1}\right)} \tag{28}$$

For the *second* set of constraints:

$$\dot{Y} = Y^{\alpha_1} x_1 M^{\delta_1} \tag{29}$$

$$\dot{V} = V^{\alpha_2} x_2 M^{\delta_1} \tag{30}$$

$$M^{*} = \left\{\frac{\delta_1}{r}\left(P Y^{\alpha_1} x_1 - w V^{\alpha_2} x_2\right)\right\}^{\left(\frac{1}{1-\delta_1}\right)} \tag{31}$$

For the *third* set of constraints:

$$\dot{Y} = x_1 Y \tag{32}$$

$$\dot{V} = V^{\alpha_2}\left(x_1 Y^{1-\alpha_1} - x_2 M^{\delta_1}\right) \tag{33}$$

$$M^{*} = \left\{\frac{\delta_1}{r} w V^{\alpha_2} x_2\right\}^{\left(\frac{1}{1-\delta_1}\right)} \tag{34}$$

For the *fourth* set of constraints:

$$\dot{Y} = Y^{\alpha_1} x_1 M^{\delta_1} \tag{35}$$

$$\dot{V} = V\left(Y^{\alpha_1 - 1} x_1 M^{\delta_1} - x_2\right) \tag{36}$$

$$M^{*} = \left\{\frac{\delta_1}{r} x_1 Y^{\alpha_1}\left(P - w\frac{V}{Y}\right)\right\}^{\left(\frac{1}{1-\delta_1}\right)} \tag{37}$$

In addition, an expression for equilibrium growth and equilibrium change of variable factor demand can be derived for each model, by substituting equilibrium investment into the constraint equations. We then get the equilibrium development of the firm as function in terms of its state variables *Y* and *V* only.

Consider first models 3 and 4. The basic hypothesis is that firms aim at a certain growth in output and a certain growth in productivity respectively. If it can be demonstrated that growth of output differs systematically over firms, model 3 can be rejected; if it can be shown that productivity growth varies systematically over fifms, this invalidates model 4. We test for systematic variation in output growth (model 3) as follows. First the data are ordered according to some criterion, e.g. according to size Y. Then we regress output growth on a vector of constants. If output growth would vary systematically with the criterion variable, here size Y, we are bound to find first order autocorrelation. By testing for autocorrelation, we thus test for systematic variation in the data. This test has been performed with two data sets for size: output represented by Y_1 and Y_2. Variable factor inputs were represented by V_0, V_1 and V_2 respectively. Data have been ordered according to seven different criteria: size Y_i, with $i = 1,2$, three measures of productivity Y_i/V_j, with $j = 0,1,2$, investment divided by size M/Y_i, average investment lag (as defined in section 5 of chapter 4), and an indicator derived from the number of years of use of front office automation and back office automation. Together this gives 14 regressions to be estimated, and 14 Durbin-Watson statistics as tests on first order autocorrelation. For model 3, the outcomes are reported in Table 4.

Table 4: Results of tests for structure in output growth.

	ordering	*coeff.*	*st.err.*	*sign.*	*D-W-stat.*
$\dot{Y}_1/Y_1$	Y_1	.071	.0025		1.904
	Y_1/V_0				1.729
143 obs.	Y_1/V_1				1.918
	Y_1/V_2				2.057
	M/Y_1				1.841
	inv.lag				1.972
	b.&f.o.aut.				1.674
$\dot{Y}_2/Y_2$	Y_2	.046	.0019		1.925
	Y_2/V_0				1.842
136 obs.	Y_2/V_1			*	1.565
	Y_2/V_2				1.845
	M/Y_2			**	1.479
	inv.lag				1.737
	b.&f.o.aut.				1.840

Significance points at the 5% level for the Durbin-Watson statistic, when the number of observations is more than 100, and there is one regressor are: $d_L = 1.65$, $d_U = 1.69$. Significance points at the 1% level are: $d_L = 1.52$, $d_U = 1.56$. The D-W statistic is insignificant in all cases where production capacity is measured by mutations of current accounts Y_1. Thus the growth in the number of mutations in current accounts does not seem to be related to size, productivity or technological level. Maybe this proxy for production capacity is too much determined by fluctuations in effective demand, and suffers from the fact that in banking the utilization rate of the capacity to produce account mutations is not likely to be constant. When capacity is measured by the number of current accounts Y_2, the D-W statistic is significant in two instances. It seems that productivity could influence growth, although the evidence is not very strong, but even more so that investment per unit of capacity correlates with growth of output, a feature present in models 1, 2 and 4, but not in model 3. More investment seems to go together with higher growth rates. All together it appears that the evidence on the basis of which model 3 could be rejected is rather weak and not consistent over different measures of output. However, there is an indication that relative expansion correlates with investment per unit of output, which in principle should suffice to reject model 3.[1]

Model 4 can be tested in a similar way. Again we order the data according to mounting output Y, labour productivity Y/V, investment per unit of output M/Y, investment lag, and years of use of automation equipment. Using two measures for Y and three for V, we arrive at six data sets to evaluate the model. The results are reported in Table 5.

The evidence of systematic variation in the data, although not entirely consistent either, is more convincing here than in the case of model 3. Again stronger results surface when numbers of current accounts Y_2 are used as capacity measure, then when mutations in current accounts Y_1 are used. Furthermore, it seems that changes in productivity correlate most with investments per unit of output, which would be in accordance with models 1 to 3, and with years of use of back and front office automation equipment. Strangely, we see more systematic variation, if we consider total labour costs than if narrower measures for costs of labour are considered. The evidence against model 4 seems to constitute some ground for rejecting it, more consistent than the evidence that there is against model 3.

This brings us to model two. To test this model against a more general model, like e.g. model 1, we proceed in a similar vein as above. First add random terms ε_1 to the equations for $\dot{Y}$ and ε_2 to the equations for $\dot{V}$ above. Then take, for model 1 and 2, the ratio of these two equations. We get expressions of the following general type:

$$\frac{\dot{Y}}{\dot{V}} = \frac{Y^{\alpha_1}}{V^{\alpha_2}} F\left(\frac{Y}{V}, M, \varepsilon_1, \varepsilon_2\right) \tag{38}$$

In case model 1 is appropriate, the function F is a complicated function of productivity (and other variables); in case model 2 is right, F does not vary with productivity. The data have been ordered according to increasing productivity Y/V. The function F has been replaced by a constant and the equation has been estimated. If autocorrelation would be found, this would indicate that the disturbances correlate with productivity Y/V, which would necessitate us to reject model two. Estimation of equation (38) as it stands, using non-linear methods, proved impossible due to severe multicollinearity of output Y and labour demand V. Taking the subsample of banks for which both $\dot{Y}_t$ and $\dot{V}_t$ are positive, equation (38) has been estimated in loglinear terms, using ordinary least squares. The fact that a non-random subsample is used for the regressions does not invalidate the test, unless there would be reasons to assume that banks that decrease labour demand would be structurally different from those that increase labour demand. If it can be

[1] The coming across one single black swan is sufficient proof against the claim that all swans are white, on the condition that the perception that this swan is of black colour does not depend on the disputable quality of ones spectacles.

Table 5: Results of tests for structure in productivity growth.

	ordering	*coeff.*	*st.err.*	*sign.*	*D-W-stat.*
$\dot{Y}_1/Y_1 - \dot{V}_0/V_0$	Y_1	.062	.0044	*	1.532
	Y_1/V_0				1.676
143 obs.	M/Y_1			**	1.292
	inv.lag				1.649
	b.&f.o.aut.			**	1.243
$\dot{Y}_1/Y_1 - \dot{V}_1/V_1$	Y_1	.083	.0071		1.767
	Y_1/V_1			*	1.626
143 obs.	M/Y_1			*	1.626
	inv.lag				1.846
	b.&f.o.aut.				1.672
$\dot{Y}_1/Y_1 - \dot{V}_2/V_2$	Y_1	.041	.0085		1.942
	Y_1/V_2				2.007
143 obs.	M/Y_1				1.879
	inv.lag				2.074
	b.&f.o.aut.				1.985
$\dot{Y}_2/Y_2 - \dot{V}_0/V_0$	Y_2	.037	.0040	*	1.557
	Y_2/V_0			**	1.497
136 obs.	M/Y_2			**	1.461
	inv.lag			*	1.597
	b.&f.o.aut.			**	1.147
$\dot{Y}_2/Y_2 - \dot{V}_1/V_1$	Y_2	.057	.0071		1.858
	Y_2/V_1			**	1.287
136 obs.	M/Y_2			*	1.580
	inv.lag				1.820
	b.&f.o.aut.			*	1.577
$\dot{Y}_2/Y_2 - \dot{V}_2/V_2$	Y_2	.017	.0087		1.927
	Y_2/V_2				1.985
136 obs.	M/Y_2				1.871
	inv.lag				2.079
	b.&f.o.aut.			*	1.591

proved for a restricted sample that model 2 does not represent the data adequately, it follows that the model will also be unable to represent the full data set adequately. The results of the exercise are listed in Table 6.

The standard deviations in the estimates of the α's are large and none of the estimates is significantly different from one at the 5% level, nor from zero. Multicollinearity of log(V) and log(Y) could account for the fact that estimates for α_1 and α_2 are rather similar in every regression, but vary quite substantially over the regressions. All Durbin-Watson statistics clearly point in the direction of first order autocorrelation. The ratio of the change in output and the change in

Table 6: Results of estimation of equation (38).

		α_1	α_2	*const.*	*obs.*	*D-W*
Y_1	V_0	1.846	-1.229	-8.955	75	1.003
		(.836)	(.788)	(6.061)		
	V_1	.514	-.616	1.956	56	.768
		(.834)	(.874)	(4.074)		
	V_2	1.314	-1.324	.1465	88	.870
		(.787)	(.793)	(3.009)		
Y_2	V_0	.035	.211	-.364	72	1.178
		(.719)	(.694)	(2.272)		
	V_1	-.660	.675	-8.464	52	.898
		(1.241)	(1.279)	(7.841)		
	V_2	-.012	-.186	-2.343	84	.956
		(.996)	(.997)	(5.270)		

Standard errors in parentheses.

labour demand, and thus according to the model the direction of technical change, is positively correlated with the ratio of output and labour demand. We can thus conclude that model 2 does not provide an accurate representation of technical change in banking, as reflected in the data.

So far we have considered three models, in which the direction of technical progress was restricted in some specific way: model 2, in which there is a fixed direction of technical progress, model 3, in which there is a fixed output growth, and model 4, in which labour productivity growth is fixed. Assessing the relative merits of these models, one can conclude on the basis of the evidence collected that the case against model 2 is strongest. It seems that banks of different productivity invest in different directions of technical change. The evidence against model 4 is less strong, but still valid: productivity growth varies over banks, both with present levels of productivity and with current technological levels. The case against model 3 is weakest, but this could be a consequence of the disparity between the capacity growth which is modelled and the growth of output which is measured. Thus rejecting models 2, 3 and 4, we continue with the more general model 1, according to which the direction of progress can vary over firms.

We can now return to our question about the similarity of decisions concerning the instruments of progress β and *M* over all banks. The overall conclusion is, first of all, that banks of different size choose different amounts of investment, but that no relationship between productivity and amount of investment has been detected, and, secondly, that banks do not seem to choose the same direction of technical progress $g(\beta)/\beta$, in particular, that banks with different productivity choose different directions of technical progress.

5.3.2 Same constraints, different decisions

The preliminary conclusion of the analysis so far is that the observed variation in β_j and M_j is not random, but systematically dependent on firm characteristics. The hypotheses that there is a fixed growth of productivity, that there is a fixed growth of capacity, and that there is a fixed (technologically predetermined) direction of technological change could all be rejected on empirical grounds. It seems that banks' choices of values for their instruments β and *M* are determined by firm characteristics, and that they therefore attain different rates of expansion and productivity growth. Now assume that there is a common investment function $h(M)$ and a common trade off function $g(\beta)$. If banks choose different investment amounts and different directions of progress, they choose different points on $h(M)$ and on $g(\beta)$. If this assumption is right, then the data must be able to tell us more about the parameters of these functions $h(M)$ and $g(\beta)$.

In this section we start out from cell 3b in Table 2: let us assume that there is one common function $g(\beta)$ which is the same for every bank, and likewise that there is one common function $h(M, \tau_1 \ldots \tau_n)$ which expresses that the rate of progress not only depends on investment, but also on the technological level, indicated by technology proxies $\tau_1 \ldots \tau_n$ of the bank. Assume, like above in model 1, that $g(\beta)$ is downward sloping and concave, and can be approximated by a second degree polynomial, and that $h(M, \tau_1 \ldots \tau_n)$, which must be sloping upward and concave for given values of τ_i, is well represented by $M^{\delta_1} \cdot \prod_{i=2}^{n} \tau_i^{\delta_i}$.[2]

Three variables have been used to represent the technology variable: the number of years that a firm disposes over back office automation equipment, the number of years it disposes over front office equipment, and the average number of years back that the firm invested in new technology. Table 7 gives correlation coefficients between these indicators and investment, capacity and productivity.

Table 7: Coefficients of correlation between indicators of the technological level of banks, investment, size and productivity.

	cap.	*prod.*	*inv/cap.*	*inv.lag*	*b.o.aut.*	*f.o.aut.*
cap. (= Y_1)	1.					
prod. (= Y_1/V_2)	.050	1.				
inv.(5yr cum.)/cap.	-.005	-.292	1.			
inv.lag	.069	.254	-.240	1.		
years b.o.aut.	.588	.123	.273	.431	1.	
years f.o.aut.	.528	.095	.498	.128	.592	1.

The three indicators for technological level are only weakly correlated. It is remarkable that the correlation coefficient between the investment lag and the number of years that front office equipment has been installed is so weak, given that automating the front office entails such major expenses. The investment lag is negatively correlated with cumulated investment per unit of output, suggesting that firms that have been investing later on average have been investing larger amounts per unit of capacity. There seems to be a positive correlation between the investment lag and productivity, indicating that firms that invested longer ago are more productive, which could point at a learning effect. One should interpret this figure with care, because it does not take account of the total size of investment, only its average point in time.

[2] Using this formulation, in which the technology variables appear as a multiplicative factor, one could interpret the model also as representing either cell 2c or 3c in Table 2, saying that technological level influences the choice of direction of progress.

There is a positive correlation between the total size of investment per unit of output and the years of back office and front office automation, as could be expected. Contrary to expectations, there is a negative correlation between investment per unit of output and labour productivity.

Each bank chooses a combination (β_j, M_j), given its technological level represented by its values τ_{ij}. This results in a change of output and variable factor demand. Assume that there is an additive residual term. The model for firm j now looks as follows:

$$\dot{Y} = Y^{\alpha_1}\frac{1}{2a_2}\left(1 - a_1 - \frac{P}{w}\frac{c_1}{c_2}\frac{Y^{\alpha_1}}{V^{\alpha_2}}\right)M^{\delta_1}\cdot\prod_{i=2}^{n}\tau_i^{\delta_i} + \varepsilon_1 \tag{39}$$

$$\dot{V} = V^{\alpha_2}\left(-a_0 + \frac{(1-a_1)^2}{4a_2} - \frac{1}{4a_2}\left(\frac{P}{w}\frac{c_1}{c_2}\frac{Y^{\alpha_1}}{V^{\alpha_2}}\right)^2\right)M^{\delta_1}\cdot\prod_{i=2}^{n}\tau_i^{\delta_i} + \varepsilon_2 \tag{40}$$

Note that the model is not scale invariant. This does not pose a problem here, because a change of scale of measurement of output and/or labour will not affect our parameter estimates of α_1, α_2 and δ_1. Nor will it affect the estimates of the standard deviation of the above parameters, or the sign of the parameter a_2. The effect of a change in scale would be fully absorbed by the order of magnitude of the weighted price ratio $(Pc_1)/(wc_2)$ and the parameters a_0, a_1 and a_2. Probably, a change in scale would affect the significance of the estimates of the parameters a_0, a_1 and a_2. This implies that the significance of these parameters could be manipulated, and conversely, that insignificance of the estimates of these parameters should not bother us.

The system consisting of equations (39) and (40) has been estimated with non-linear least squares methods, using both mutations in current accounts Y_1 and numbers of current accounts Y_2 as measures for output, and total labour V_0, total labour input into payment traffic V_1 and input of cashier, counter and processing work into payment traffic V_2 as proxies for variable factor demand. There are no data available for the ratio of the output price P and the price of the variable inputs w, but since the data cover only two periods, the price ratio has been assumed constant over time. It was not always possible to estimate the ratio $(Pc_1)/(wc_2)$ as a parameter. Using the data Y_1 in combination with V_2, and Y_2 together with V_0, estimates of the price ratio were obtained, but for the combination of Y_1 and V_1 this turned out impossible, due to singularity of the data. For Y_2 together with V_1 and V_2, the estimates of the other parameters turned out to be insensitive to variation in a-priori values for the price ratio. Therefore $(Pc_1)/(wc_2)$ has been set to one in all but two estimations. Moreover, it turned out to be impossible to find least squares solutions for the variant where capacity is Y_1 and labour demand is V_0; the estimation results appear to be rather unstable, probably also due to multicollinearity of the independent variables.

The results of the estimation procedures are reported in Tables 8 and 9. In the first table there are results of four estimations using Y_1 and V_2, and four estimations using Y_2 and V_2 respectively. In the first estimation, no technology variable τ was included (which would correspond to cell

2b in Table 2). In the second, τ_2, the average age of investment in automation equipment, was included. In the third, the variables τ_3 and τ_4, years of use of back office automation equipment and front office automation equipment were inserted. The fourth estimation contained all three technology variables. To get an impression of the robustness of the results when other data series are used, the Table 8 can be compared to the Table 9. In this table, similar estimations were run using the broader proxies for labour demand: V_1 and V_0. The data were ordered per period, and within every period approximately according to firm size. In Table 10, the Durbin-Watson statistics for each of the equations are listed, and also the adjusted R-squares.

Table 8: Results of estimation of system (39) and (40), using a narrow measure for labour inputs.

syst.	α_1	α_2	a_2	a_1	a_0	pc_1/wc_2	δ_1	δ_2	δ_3	δ_4
1.	.800**	.759**	-4322	398	-8.94	23.93	.244**			
	(.081)	(.099)	(15425)	(1442)	(34.90)	(40.65)	(.078)			
2.	.842**	.797**	-1689	179	-4.52	15.88	.199**	-.198		
	(.090)	(.108)	(4347)	(462)	(13.16)	(18.57)	(.088)	(.158)		
3.	.791**	.751**	-2007	121	-1.70	14.26	.287**		-.075	.000
	(.087)	(.105)	(4716)	(314)	(5.59)	(16.30)	(.092)		(.097)	(.189)
4.	.826**	.781**	-1190	113	-2.47	13.15	.210 *	-.233	-.048	.126
	(.093)	(.111)	(2688)	(272)	(7.66)	(13.83)	(.113)	(.188)	(.101)	(.210)
5.	.207**	-.109	-.861	-12.8	-14634	1.00	.150 *			
	(.065)	(.122)	(2.221)	(17.9)	(30392)		(.084)			
6.	.227**	-.081	-.330	-10.9	-22949	1.00	.098	-.192		
	(.066)	(.110)	(.809)	(14.2)	(46178)		(.097)	(.170)		
7.	.260**	-.019	-.964	-6.18	-2567	1.00	.286**		-.562**	-.146
	(.070)	(.090)	(2.086)	(7.15)	(5186)		(.102)		(.198)	(.092)
8.	.256**	-.018	-1.533	-5.82	-1557	1.00	.326**	.093	-.604**	-.161*
	(.072)	(.091)	(3.776)	(6.83)	(3517)		(.130)	(.191)	(.220)	(.096)

Standard errors in parentheses.

Legenda: system 1 to 4: data Y_1 and V_2 (143 obs.); system 5 to 8: data Y_2 and V_2 (136 obs.);

τ_2: average investment lag; τ_3: years of back office automation; τ_4: years of front office automation;

*: significant at the 10% level; ** significant at the 5% level.

First consider Table 8. The most consistent and robust result is the significance of the parameter α_1. It turns out that this parameter is significantly positive in every regression. At the same time the value turns out to be significantly below one. The order of magnitude of the estimate, however, shows some variation. If capacity is measured by mutations in current accounts Y_1, the point estimate is about three to four times as large as when number of current accounts Y_2 is used. The results for the parameter α_2 are far less robust: α_2 is significantly positive, when mutations in current accounts Y_1 is used as capacity proxy, but is insignificant, often with the wrong sign, when numbers of accounts Y_2 is used. The estimates are all significantly below one. In the cases where α_2 is significant, the point estimate for this parameter lies marginally below the estimate for α_1. This could be interpreted as an indication that the opportunities for technological change in the period which we analyse were neutral, maybe slightly biased in the direction of expansion (see section 2 above).

Another rather robust result is the significance of the parameter δ_1 which captures the effect of prior investment on capacity growth and productivity improvement. It is remarkable that the parameter estimates are consistently below .3, which indicates that the investment function is strongly concave. This would mean that there are no *static* economies of scale. On the contrary, the effect of a marginal increase in investment at any moment in time, in terms of extra capacity or growth in productivity, seems to decline rather rapidly. This does not imply anything about the existence of *dynamic* economies of scale. Dynamic economies of scale are expressed in the model through the α_i parameters. The higher those are, the stronger the intertemporal effects of investments, the more prominent the dynamic scale economies. In the case of dynamic scale economies, it is difficult to draw a conclusion on their prominence, because the point estimates of the α_i's vary quite substantially over the different data sets.

A necessary condition for a concave downward sloping function $g(\beta)$ is a negative value a_2. If a_1 is negative, then $g(\beta)$ is concave all over the permitted domain of β; if a_1 is positive, however, the second degree polynomial which we use to approximate $g(\beta)$ has a convex and a concave part. In this case, one would have to check whether the polynomial is actually concave in the region where it approximates the function $g(\beta)$. Unfortunately, of the parameters a_2, a_1 and a_0, the sign cannot be determined in any of the regressions with a 90% probability of correctness. Nevertheless, from the tables it appears that the sign of the point estimates of a_2 is always negative, and that the sign of a_1 is negative when Y_2 is the capacity proxy.

Finally there are the parameters δ_2, δ_3 and δ_4 which should capture the effect of differences in technological level on the effect of investment on growth and productivity. Evidence here is consistent as far as the signs of the parameters is concerned. Almost all point estimates of all δ_i's are negative. This indicates that firms that have invested comparatively long ago, or have had certain systems of automation already for a longer time, reap less benefits from investment than firms that have started to invest at a later date. One might interpret this as evidence that banks that have invested earlier, have gone further down the learning curve and are closer to the technological frontier, and are therefore less able to realize improvements with a given amount of investment, than banks that have started their automation program at a later date. An alternative explanation would be that prices of automation equipment have decreased to such an extent that late investors have got more value for their money, or better quality, and were thus able to realize more progress for the same amount of investment. As far as the significance of the parameters is concerned, it turns out that none of the technology parameters is significantly different from zero at the 10% level, if capacity is represented by Y_1. If capacity is approximated by Y_2, however, it seems that the number of years that a bank has an automated back office is a significant factor in the determination of present expansion and rationalization. The longer the bank uses back office automation, the less expansion and the less change in labour demand, *ceteris paribus*. It is surprising that, whereas the years of using back office automation is of significant influence, the years of using front office automation is not.

Comparing Table 8 to Table 9, it appears that there are no big changes in outcomes when using V_1 instead of V_2. However, using V_0 yields some different outcomes. Estimates for α_1 are above one, δ_1 turns out insignificant and the estimates for a_2 appear positive. The differences in estimates

Table 9: Results of estimation of system (39) and (40), using wider measures for labour inputs.

syst.	α_1	α_2	a_2	a_1	a_0	pc_1/wc_2	δ_1	δ_2	δ_3	δ_4
9.	.715**	.579**	-22.92	4.31	.630	1.00	.226**			
	(.159)	(.266)	(32.74)	(5.29)	(2.857)		(.081)			
10.	.763**	.632**	-17.89	3.80	.721	1.00	.179 *	-.194		
	(.162)	(.266)	(24.14)	(4.30)	(3.113)		(.092)	(.167)		
11.	.715**	.606**	-25.21	2.36	.289	1.00	.280**		.027	-.108
	(.161)	(.260)	(33.61)	(2.62)	(1.203)		(.096)		(.200)	(.101)
12.	.747**	.631**	-16.56	2.47	.532	1.00	.203 *	-.229	.156	-.081
	(.164)	(.258)	(22.83)	(2.76)	(2.166)		(.119)	(.198)	(.223)	(.106)
13.	.214**	-.190	-2.08	-34.89	40341	1.00	.102			
	(.058)	(.139)	(5.59)	(50.4)	(96406)		(.083)			
14.	.233**	-.166	-.80	-31.10	73560	1.00	.047	-.180		
	(.060)	(.129)	(2.08)	(44.46)	(169790)		(.097)	(.169)		
15.	.260**	-.121	-1.58	-22.31	18318	1.00	.188 *		-.552**	-.083
	(.060)	(.118)	(3.60)	(29.7)	(44031)		(.102)		(.197)	(.093)
16.	.257**	-.119	-1.99	21.27	13658	1.00	.210	.047	-.572**	-.091
	(.061)	(.119)	(5.08)	(28.4)	(35649)		(.133)	(.190)	(.218)	(.097)
17.	1.179**	.447**	550738	-3449	5.05	.573**	.101			
	(.099)	(.208)	(1306952)	(7274)	(10.76)	(.135)	(.073)			
18.	1.239**	.457**	392789	-6316	23.99	.603**	.010	-.299**		
	(.102)	(.225)	(969077)	(14166)	(56.63)	(.141)	(.085)	(.147)		
19.	1.265**	.449 *	1209313	-8223	13.21	.566**	.110		.014	-.535**
	(.108)	(.232)	(3011148)	(18682)	(30.74)	(.142)	(.093)		(.089)	(.181)
20.	1.325**	.493**	645959	-9919	36.06	.551**	.005	-.224	.067	-.438**
	(.116)	(.273)	(1647123)	(22630)	(87.90)	(.134)	(.123)	(.172)	(.099)	(.194)

Standard errors in parentheses.

Legenda: system 9 to 12: data Y_1 and V_1 (143 obs.); system 13 to 16: data Y_2 and V_1 (136 obs.);
system 17 to 20: data Y_2 and V_0 (136 obs.);
τ_2: average investment lag; τ_3: years of back office automation; τ_4: years of front office automation;
*: significant at the 10% level; **: significant at the 5% level.

of p/w in both tables are caused by differences in units of measurement of the different proxies of the capacity and variable input variables. Table 10 indicates that there could be some autocorrelation in the residuals, which would mean that the residuals would be somehow correlated with size, despite the fact that the size variable and the labour demand variable appear on the right hand side of both equations of the system. Adjusted R-squares show that the fit of the first equation of the system is considerably better than the fit of the second.

Taking all evidence together, then it may be concluded that α_1 is likely to be positive and above .2, but that α_2 may be anything between zero and one. This indicates that it is likely that there are some dynamic economies of scale, that there is technological progress and learning by doing, but that its extent is difficult to quantify. Referring back to the matrix of Table 2, it seems that we end up in the area of cells 2b, 2c, 3b and 3c. Firms make different choices on matters of size of investment M_j and on direction of progress $g(\beta_j)/\beta_j$. They do this because they have a number of different characteristics at the moment of decision making, notably size and variable factor productivity. This has been shown in the first part of this section. They also might choose differently, because the shape of the technical constraint functions $h_j(M_j)$ and $g_j(\beta_j)$ vary with

Table 10: Some statistics from estimation of equations (39) and (40).

system	*obs.*	*adj.*R^2 *(eq.* $\dot{Y}$*)*	*adj.*R^2 *(eq.* $\dot{V}$*)*	*DW (eq.* $\dot{Y}$*)*	*DW (eq.* $\dot{V}$*)*
1.	143	.785	.191	1.594	1.712
2.	"	.785	.196	1.592	1.722
3.	"	.785	.192	1.638	1.708
4.	"	.785	.196	1.600	1.722
5.	136	.735	.043	1.373	1.534
6.	"	.737	.017	1.382	1.532
7.	"	.755	.015	1.468	1.531
8.	"	.755	.020	1.471	1.532
9.	143	.772	.243	1.509	1.355
10.	"	.772	.249	1.500	1.354
11.	"	.771	.240	1.563	1.355
12.	"	.771	.248	1.514	1.355
13.	"	.746	.032	1.426	1.531
14.	"	.748	.011	1.437	1.532
15.	"	.762	.008	1.499	1.536
16.	"	.762	.008	1.500	1.537
17.	136	.756	.204	1.475	1.178
18.	"	.760	.211	1.533	1.181
19.	"	.776	.201	1.531	1.171
20.	"	.774	.206	1.544	1.178

the technological level τ_j of the firm. Evidence here is less strong, but it seems that those banks that embarked earlier on the path of automation get less returns from their present investments than firms that moved later. This would re-assert claims that catching-up goes with lower costs than moving first, and that for firms operating at the technological frontier progress is more costly. Concerning the shape of the constraint functions, we found support for the assumption that $g(\beta)$ is concave. Furthermore, we saw that $h(M)$ is likely to display strongly decreasing returns to investment at any moment in time, and thus that there is no indication of static economies of scale of investment.

The aspects of the model that have been tested so far are mainly concerned with the validity of the technical constraints. A complementary way to test the model, especially with respect to its assumptions on the rationality of investment behaviour, would be to estimate equation (28) above, which expresses the optimal amount of investment for a firm of size Y_j and productivity Y_j/V_j. In doing that, we not only test the validity of the technical constraints, but also the rationality of the investment choices of firms. This has been attempted, using data for Y and V for 1985, and cumulated data for investment in automation equipment M from 1985 to 1987. Equation (28) has been estimated using two different search algorithms for nonlinear models, the so called direct search algorithm and the complex method, which both attempt to find a set of parameters which minimizes the sum of squared residuals in a restricted parameter space.[3] Neither method converged to a global optimum within the space of admitted parameter values. It turned out that changes in the value of the objective function are minor over large ranges of the parameters.

In principle, the equations for $\dot{Y}$, $\dot{V}$ and M^* could also be estimated as one system of three equations. Because this would involve some complicated estimation procedures, but also because the least squares criterion does not yield a clear estimate for the parameters of equation (28) and the objective function seems to find its minimum on a plain in the parameter space, this simultaneous estimation has not been attempted. If the squared residuals of equation (28) are minimized by a large variety of parameters, the outcomes of the simultaneous estimation of equations (26), (27) and (28) are bound to be the same as the outcomes of the simultaneous estimation of only equations (26) and (27), as reported in Table 8 and 9.

Finally we may return to the hypotheses presented in Table 1 above. These concern differences in direction and amount of investment, for firms of different size and variable factor productivity, given common functions $g(\beta)$ and $h(M)$. From equations (2) and (3), using different measures Y and V, the direction of technical progress of each firm can be calculated as $g(\beta)/\beta = 1 - \left(\dot{V} Y^{\alpha_1} c_1\right)\!/\!\left(\dot{Y} V^{\alpha_2} c_2\right)$. Estimates for α_1 and α_2 were taken from The ratio c_1/c_2 was set equal to unity, as this only implies a monotonous transformation of the series. Tables 8 and 9 for the various sets of data. The bottom left hand corner of Table 1 implies that $g(\beta)/\beta$ should be lower for more productive firms, *ceteris paribus*. The top left hand corner implies that $g(\beta)/\beta$ should be lower for larger firms, if $\alpha_1 > \alpha_2$. Regressions were run of $g(\beta)/\beta$ on size Y and productivity Y/V. Some weak evidence could be found for the first hypothesis (see Table 11): $g(\beta)/\beta$ seems to vary negatively with productivity. Coefficients are consistently negative, but only significant in two cases. No significant relationship between $g(\beta)/\beta$ and size was found, an outcome not unexpected in cases where $\alpha_1 = \alpha_2$.

Table 11: Results of regressing $g(\beta)/\beta$ on size Y and productivity Y/V.

		Y	Y/V	*const.*
Y_1	V_2	-.000	-2.214	3.340
	V_1	(.000)	(1.487)	(2.239)
	V_2	-.000	-4.803	15.52
Y_2	V_1	(.000)	(14.736)	(15.69)
	V_0	.014	-.000**	5431**
		(.055)	(.000)	(1996)
		.048	-.000**	7329**
		(.061)	(.000)	(1943)
		.006**	-22.95	30.09
		(.002)	(17.50)	(51.88)

Standard errors in parentheses;
*: significant at the 10% level; ** significant at the 5% level.

[3] For a description of these search algorithms, see e.g.: B.D. Bunday and G.R. Garside, Optimization Methods in Pascal, 1987.

The top right hand corner of Table 1 says that larger firms should invest more and the bottom right hand corner says that more productive firms should invest more, provided their direction of progress $g(\beta)/\beta < 1$. The high correlation between investment and size has already been mentioned. To test the latter hypothesis, the sample of banks was divided in two parts, a group of banks for which $g(\beta)/\beta < 1 \Leftrightarrow dv > 0$, and a group for which the converse was measured.[4] Then a positive relationship between variable factor productivity and investment was searched for the first group and a negative for the second. Here no evidence supporting the hypothesis could be found. All together it seems that the relationship between size and variable factor productivity of firms and their investment choices is weak. This lack of regularity casts some doubt on the assumption that investment opportunities can be characterized by functions $g(\beta)$ and $h(M)$ which are common to all banks in the sample.

5.4 Conclusion

In this chapter, a number of aspects of the model developed in chapter 3 were put to the test, using data from a banking organization. The model describes the development of capacity and productivity of a firm over a longer period of time. A more comprehensive test of the model would thus require time series data. Since there are only short time series in the data base, but extensive cross section information is available, only a myopic version of the model could be estimated. The emphasis of the testing procedures is not on the validity of the model *per se*, but on the extent to which firms are *similar* in the technical constraints they face and consequently in their investment behaviour. We try to determine, to what extent the firms in the data set are in a similar situation and take the same decisions, and to what extent they follow each other on a comparable path, where there are first movers, followers and laggards.

There are four key variables in the model: production capacity, variable factor productivity, investment and technological level. Output series were used as proxies for production capacity, labour input series were used for variable factors of production, and five years cumulated investments in automation equipment was taken for investment. The technological level was approximated by three different indicators, the average lag in investments in automation equipment and the number of years of use of back office and of front office automation equipment respectively. Each of the main data series selected suffers from some drawback: the proxy for production capacity does not take fluctuations in capacity utilization into account; the proxy for variable inputs is in fact largely fixed, and the proxy for investments has a rather arbitrary lag structure and does not include investments in human capital and other efficiency improvement. The estimation results thus have to be considered with some restraint.

One of the main questions addressed was, whether expansion of output capacity and raising of productivity are processes in which there is a cumulative element or not: do large banks grow by the same or by a larger volume per unit of investment than small banks. Large banks in our data set commonly expand more in absolute terms than small banks, but they also tend to invest more: absolute growth, size and investment are highly correlated. This is not surprising, but the question remains *why* large banks grow faster in absolute terms: because they invest more, or because their size lets them reap more advantage from an amount of investment, though the

[4] Since $g(\beta)/\beta = 1-\left(\dot{V}Y^{\alpha_1}c_1\right)/\left(\dot{Y}V^{\alpha_2}c_2\right) < 1 \Leftrightarrow (\dot{V}Y^{\alpha_1})/(\dot{Y}V^{\alpha_2}) > 0$ and $\dot{Y}, Y^{\alpha_1}, V^{\alpha_2} > 0$, whether $g(\beta)/\beta$ exceeds one or not depends on the sign of $\dot{V}$ and is independent of α_1 and α_2.

marginal return to investment decreases quickly. Investment may either lead to more efficient production technology which can better be exploited by larger firms, which is why they invest more despite rapidly decreasing returns to investment (in a static sense), or it may only lead to expansion of the capital stock, a process where decreasing returns to investment may be less rapid and which may be exploited by large and small firms alike. The estimation procedures that have been used try to disentangle the influence of investment and the influence of size. The results indicate that, on the one hand, capacity increase depends on investment, but that there are strongly decreasing returns to investment. On the other hand, the effect of investment on capacity increase seems to be related to current production capacity, and thus there is likely to be a cumulative effect. Therefore, the presumption that investment is into production technology which brings its largest benefits to large firms cannot be refuted. A significant effect of the current volume of demand for variable inputs on growth of variable factor demand, however, is not consistently found in the data. If there is a positive effect here, parameter estimates indicate that it may be smaller than the effect of production capacity on capacity growth.

The second topic in this chapter was the question to which extent firms follow the same track of expansion and productivity increase. On the one hand we considered whether the trade-off between capacity expansion and improvement of variable factor productivity can be represented by one single function $g(\beta)$ for all firms in the sample, and whether firms tend to make the same choice concerning the location on this curve. On the other hand we considered whether the effect of investment on capacity and productivity can be represented by one single function $h(M)$ or $h(M,\tau)$ for all firms in the sample, and whether firms are likely to choose the same amount of investment. Some arguments were developed, why firms of different size or variable factor productivity would opt for different courses of investment.

The data reveal a large variety in choices by different firms. Our analysis showed that these differences are not random. It was apparent that differences in amount of M investment are primarily related to differences in size, but it could not be shown that they also vary with variable factor productivity. Concerning the direction of investment, four alternative hypotheses were considered: a constant rate of expansion, a constant rate of productivity increase of variable factors, a fixed direction of technical change $g(\beta)/\beta$, and a quadratic approximation of the function $g(\beta)$. The first three alternatives, all implying that firms in some sense invest in the same direction and that deviations from this course are random, were refuted by the data. The fourth alternative, that the trade-off function $g(\beta)$ can be approximated by a parabolic function, was retained as working hypothesis. Finally, we considered the possibility that expansion and rationalization depend on the technological level of the firm. Only weak evidence of such a relationship could be found.

Accepting the hypothesis that firms make different decisions on direction and size of investments in technical progress, we turned back to the hypotheses of Table 1, to see whether the differences in choices were in accordance with the patterns derived theoretically. Evidence to support the theoretical claims is weak. Large firms invest more than small firms, but the direction of investment seems to be unrelated to size. There is an indication that more productive firms invest relatively more in expansion, but evidence remains inconclusive. No relationship between variable factor productivity and size of investment was found. This implies that the assumption that all firms in the sample face similar functions $g(\beta)$ and $h(M)$ is not confirmed.

The results should be judged against the background of the limitations of both data and model. The conceptual problems with the proxies for model variables and the lack of time series data have been mentioned already. Given the limited scope of the model, the results are not unsatisfactory. The model abstracts from product development, does not account explicitly for spill-over effects of adoption elsewhere and for bandwagon effects. It abstracts from certain types of strategic behaviour and expectation formation mechanisms, and does not allow for any discontinuities, nor in investment, nor in expansion or change of productivity, factors which are all likely to be important in the case of the banking industry. Possibly the model therefore puts undue stress on arguments of marginal costs and benefits. Banks operate on markets characterized by monopolistic competition or oligopoly. Strategic considerations are likely to be important in decisions on investment. Also, because of uncertainties connected to the rapid speed of technological development in information technology, bandwagon effects are likely to be prominent. On the other hand, the price of banking products like a savings account and a payment transfer, as well as its costs, are generally obscure to the customer. The demand for these products is therefore not a function of the price. Thus a number of factors that might be of influence in banking do not appear in the model, whereas a mechanism that does appear in the model, price sensitivity of output, is likely to be weak in banking. Nevertheless, although the model probably does not capture a number of mechanisms which influence investment planning within a banking firm, especially in the short term, it does seem to capture some part of the trend.

6. Diffusion models; theory and empirics

6.1 Introduction

So far we approached the problem of explaining technological change in the banks of our database from the viewpoint of the profit maximizing firm. The models above are a representation of rational firm behaviour, assuming profit maximization and opportunities for technical progress. Although these models may serve to illuminate some aspects of the nature of investment planning of firms, they turned out to be of limited value in explaining changes in growth and productivity in banks. We saw that data for size and investments in automation equipment are strongly correlated, but that labour productivity follows a rather irregular pattern over firms, showing a weak relation to size and investment. Also, nor investments in automation, nor the use of certain systems, seemed to bear a firm relationship to growth of production or productivity.

However, whereas we see irregular patterns of industry development when we consider firm level quantitative variables like productivity, we find fairly regular patterns when we look at aggregate level qualitative measures as whether banks use a certain automation systems in a specific year or not (see Figures 4 and 5 in chapter 4). Although it turns out difficult to explain technological change with a model of optimal investment of firms, there is an orderly diffusion phenomenon discernible at the aggregate level. Although the links between individual adoptions of technology and their supposed effects on productivity are not easily apparent in the data, the number of adoptions of a system in a specific year seems to develop regularly. Thus we observe irregularity at the micro level, measuring productivity, but regularity at the aggregate level, counting adoptions. Moreover, the shape of this aggregate level regularity is the familiar sigmoid curve, a rather robust finding in diffusion research. In this chapter we shall concentrate on possible explanations of this industry level pattern of development.

One approach might be to aggregate microeconomic models of firm behaviour. However, it should be noted that, even if a microeconomic model explains firm behaviour accurately, additional assumptions are needed to arrive at an explanation of aggregate phenomena. The firm model used above provides for a slow approach of the newest technology by firms that are not at the technological frontier, which makes a gradual diffusion of technologies at the industry level likely. However, the shape of the diffusion path is indeterminate so far, because it depends on the conditions of firms at the start of the process. Aggregate phenomena can only be explained with these type of models, when some distribution over characteristics that differentiate between firms in the industry at the time diffusion starts is specified. Following this course, we would end up with some type of probit model, where the difference between firms would be characterized by a distribution over size, over current variable factor productivity, or both; it can also be over technological capabilities, price expectations, or other. This distribution determines the shape of the diffusion curve.

The distribution of a characteristic over firms would have to be added to the existing firm model. This prompts three questions. First of all, one may want to know what explains the distribution. Secondly, one may wonder whether this distribution stays stable over the course of the diffusion process; introduction of technology may change characteristics of firms like size, market share, productivity, etc. Thirdly, it is doubtful that a model that's outcome depends completely on the specification of a distribution is robust in its predictions.[1]

Moreover, aggregating micro-models is a promising approach, when one can be convinced of the explanatory power of the microeconomic base. Because it was argued in the foregoing chapter that the microeconomic model available could only explain the evidence to a limited extent, it may be interesting to supplement the analysis attempted there with an alternative approach. In this chapter, we shall try to explain and model adoption, the patterns of consecutive adoptions in Figures 4 and 5 of chapter 4, at the aggregate level.

There are two cases when an aggregate approach may be superior for practical purposes to aggregating microeconomic models. The first is, if the decision to adopt is predominantly a bandwagon effect and if there are important technology spill-overs from adoption. The probability of adopting is then to a large extent determined by other adoptions in the industry, and diffusion takes on the character of a coherent process at the aggregate level. There are different reasons why spill-overs may be of substantial influence for technology adoptions. Earlier adoptions can influence later adopters through a spread of information and skills. This can take place through communication, but also through labour mobility. There can be effects of adoptions on price levels, both on input and output markets. These can occur, because firms improve their productivity and expand their capacity. There can be network externalities. All of these three effects, information spill-overs, price effects and network externalities, may influence the adoption decision of a next potential adopter.

Secondly, an aggregate model might be appropriate if regularity in fact only exists at the aggregate level. If there is a variety of behaviour at the micro level, but a mechanism of selection at the aggregate level that determines which firms grow and which decline, then a variety of types of firm behaviour may result in similar aggregate features. Different firms may decide according to different deterministic or stochastic rules, but competition on the market, which functions as a mechanism selecting successful and more productive enterprises, imposes certain regular patterns at the level of the industry. Competition may produce aggregate regularity out of underlying variety: there is a chance mechanism operative that gives the relatively more efficient firms a higher probability of survival and growth than the relatively less efficient firms, no matter how these efficiency differentials arose. In this way, a mechanism at the aggregate level, instead of a distribution over characteristics at the micro level, accounts for order at the

[1] As Dosi writes: "[..] the 'rational-equilibrium' approach loses interpretative significance the more the diffusion process is influenced by particular distributions of the expectational and technological characteristics of agents. In this sense, 'equilibrium approaches' show the same limitations, and more so, as so-called rational expectation models in macroeconomics: 'equilibrium paths' - whenever they exist - are not independent of the distribution of beliefs, technological capabilities and learning processes of individual agents. In fact, one may simply check the robustness of a unique equilibrium diffusion path [..] allowing for different stochastic disturbances on, e.g. expectations: in general, one cannot presume the equilibrium path to be even *locally* stable; hence also the conclusions based on the properties of 'perfect' equilibrium diffusion processes cannot be presumed to hold." (Dosi, 1991, pp.200-201).

aggregate level and the regularity at the aggregate level is relatively *independent* of the decision rules or routines that agents follow at the micro level. This could help to explain the robustness of the sigmoid curve as a description of different types of diffusion phenomena.

The epidemic diffusion model is a model describing adoptions at the aggregate level. It is based on mechanisms of information spill-overs which are endogenous to the diffusion process itself. This type of model will be our starting point below. It will be used to describe the spread of three innovative techniques through our sample of banks. Chapter 3 above presented a model describing the development of a firm that introduces ever more efficient technology over time. In chapter 5, we used this model to consider a sample of banks developing parallel to each other. The model below will describe the development of the use of a technique through time. Then we shall use this model to consider a sequence of techniques that have been introduced to our sample of banks. We do not model the (rational) decision making of firms, but only assume that adoptions of new technology are mainly caused by earlier adoptions in the industry. Typically, we assume that new technology is surrounded by uncertainty, and bandwagon effects are dominant in technology adoption decisions.

6.2 Diffusion and probabilities of transition

The adoption of an innovation is a qualitative change in technique. As an innovation diffuses, it often undergoes major improvements in terms of efficiency, as a consequence of learning by doing and learning by using. This explains the observation that the differences in terms of efficiency between an old technique and an innovation are often smaller than between qualitatively similar techniques, where the newer variant is an incremental improvement over the older variant [see Rosenberg (1976), Freeman (1988)]. Thus there is often a continuum of techniques in operation that have been developed subsequently, where the qualitative gaps between some techniques are larger than between others, but where large qualitative differences are not always parallelled by large differences in efficiency.

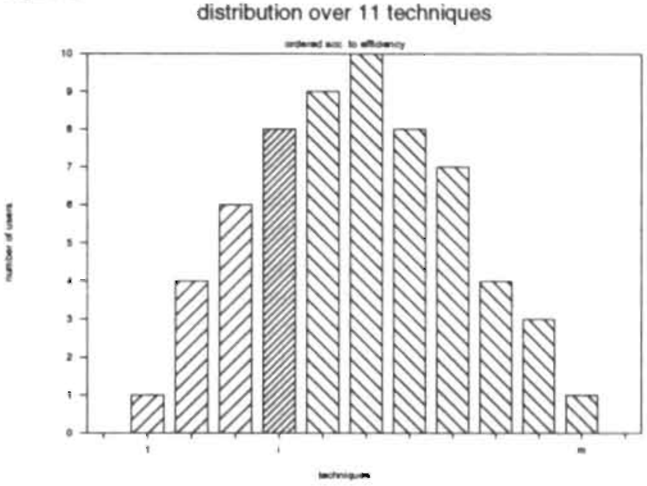
In the range of available techniques, there are obsolete techniques, older established techniques, average practice techniques, newer techniques, best practice techniques and experimental techniques, all with their specific yields, productivities, risks and market prices. Firms choose to work with different techniques according to their capabilities, risk perception and experiences. The newer the technique, the higher the requirements of technological capabilities, skills and knowledge to the using firm. Also, characteristics of techniques may be different in terms of flexibility and quality of the output. Markets for different techniques may be different. Whereas older techniques may be spread through competitive markets, newer techniques may be offered by oligopolists or a monopolist.

Assume that there are different techniques which can be numbered from 1 to m in order of increasing efficiency at current prices. Assume that every firm makes use of one technique at any period in time. At any moment in time t there is a distribution of firms over the m techniques. Figure 1 pictures a hypothetical distribution of users over techniques 1 to m. The change over time of the distribution determines the diffusion pattern of the different techniques. Two different models that describe the change of such a distribution have been developed by Iwai (1984a&b) and by Metcalfe (1988). Denote the number of firms using technique j by n_j, where $1 \leq j \leq m$.

Let p_{ji} indicate the probability that at some moment *t* firms using technique *j* decide to switch to technique *i*. Assume that there are no firms entering or exiting the industry. Diffusion of technique *i* can now be expressed as:

$$\dot{n}_i = \sum_{j=1}^{i-1} n_j p_{ji} - n_i \sum_{j=i+1}^{m} p_{ij} \qquad (1)$$

Figure 1



Time indices have been dropped. Equation (1) says that the change in the number of firms that uses technique *i* is the sum of the number of firms that used a less efficient technique *j* before $(1 \leq j \leq i - 1)$, and decided to switch to technique *i*, minus the number of firms that used technique *i* so far, but decided to move on to a more efficient technique *j* $(i + 1 \leq j \leq m)$. To account for the possibility of entry and exit, the model has to be expanded by specifying probabilities of entering ar any technological level, and of exiting the industry from any level. This is not elaborated here, since in the empirical case under consideration, there is exit nor entry. Diffusion patterns now depend on the specification of the probability p_{ji}. There is an abundance of possibilities here, all lying between two extremes. One extreme would be $p_{i,i+1} = z$, with *z* going to zero, which describes the slowest possible diffusion, where every firm goes step by step through all the techniques *1* to *m*. This specification would be adequate, if diffusion were a type of learning process, where a firm goes gradually down the learning curve and where no intermediate stage can be skipped. The other extreme would be $p_{im} = z$, with *z* going to 1, which describes immediate diffusion of the best practice technique. This specification would be adequate, in case there is no need to bother with any of the steps in between, because the best practice technique is clearly superior to everything else, for every potential adopter. This could be if there is a large cost advantage, the risks are low, and no special capabilities to employ this technique are required.

We shall consider a few intermediate possibilities. Suppose that the chance that a firm goes from j to i can be separated into two components, the chance that a firm leaves j and the chance that it goes to i, given that it has left j: $p_{ji} = p_j \cdot p_{i|j}$. We consider two different hypothetical mechanisms of diffusion and for each two alternative specifications. The first is a push mechanism and the second is a pull mechanism. The difference between the two is the reason why firms decide to switch from one technique to another.

The economic viability of a technique is a function of its relative efficiency and prices of inputs and outputs. These depend not only on the technique of any single firm, but also on the techniques used by other firms in the industry. Thus the whole distribution of firms over techniques, as pictured in Figure 1, determines whether a technique is attractive to operate or not. The impulse on any firm to switch to a better technique is a function of all other techniques being used. Firms can either be pushed out of an old technique, or they can be pulled towards a new technique.

A push mechanism can be said to operate, if the chance that firms decide to leave a technique j, indicated by p_j, depends on the 'pressure' on this technique. The pressure on a technique is a positive function of the share of firms that work with a more efficient technique. As a larger share of firms produces more efficiently, chances that prices are competed down rise, profits tend to erode, and pressure to switch to a more efficient technique mounts. We consider two specifications for p_j, one where the firm that operates technique i takes all other firms into account, the other where it looks ahead at better techniques, but not back, and only regards the other firms that operate the same technique i or more efficient, up to technique m. The chance that a firm switches *from* a technique j is thus dependent on pressure. *Once the firm leaves* technique j, the chance that it goes to i, where $i > j$, is simply a function of the number of firms that already use i. The more firms use a technique, the more experience with it exists, and the smaller are the risks, the larger the network externalities, etcetera. We get the following model (1a):

$$p_{ji} = \alpha \frac{\sum_{k=j+1}^{m} n_k}{N} \cdot \frac{n_i}{\sum_{k=j+1}^{m} n_k} = \alpha \frac{n_i}{N} \qquad (2)$$

Here $N = \sum_{k=1}^{m} n_k$ is the total of firms in the industry. The first factor expresses the pressure on users of technique j, and the second the chance of going to technique i. The alternative expression (1b) would be:

$$p_{ji} = \alpha \frac{\sum_{k=j+1}^{m} n_k}{\sum_{k=j}^{m} n_k} \cdot \frac{n_i}{\sum_{k=j+1}^{m} n_k} = \alpha \frac{n_i}{\sum_{k=j}^{m} n_k} \qquad (3)$$

The last factor is the same as in equation (2), and the first factor has a more restricted denominator. These cases can be contrasted with cases where a pull mechanism can be said to operate. Here the chance that firms coming from technique *j* decide to go to a specific technique *i*, indicated by $p_{i|j}$, depends on the 'attraction' of this technique. The attraction of a technique is the mirror image of the pressure on a technique. It is a positive function of the share of firms that work with a less efficient technique. As a larger share of firms produces less efficiently, chances that one can secure a higher profitability with such a technique rise, and firms are pulled towards it. Here we also consider two specifications for $p_{i|j}$, one where the pull of technique *i* is determined by the current use of all other techniques, the other where the attraction depends only on the use of the techniques that are less efficient than *i*. The chance that a firm switches *to* a technique *i* is thus dependent on attraction. Where the firm that goes to *i* comes from is dependent on a simple chance process: every firm that operates a technique *j* that is less efficient than *i*, $j < i$, has the same chance of switching to *i*. We get the following model (2a):

$$p_{ji} = \alpha \frac{\sum_{k=1}^{i-1} n_k}{N} \cdot \frac{n_j}{\sum_{k=1}^{i-1} n_k} = \alpha \frac{n_j}{N} \tag{4}$$

The first factor expresses the pull by technique *i*, and the second the chance that an adopter of *i* came from technique j. The alternative expression (2b) would be:

$$p_{ji} = \alpha \frac{\sum_{k=1}^{i-1} n_k}{\sum_{k=1}^{i} n_k} \cdot \frac{n_j}{\sum_{k=1}^{i-1} n_k} = \alpha \frac{n_j}{\sum_{k=1}^{i} n_k} \tag{5}$$

Substitution of equations (2) to (5) respectively in equation (1) yields four different diffusion models:

$$\dot{n}_i = \alpha \frac{n_i}{N} \left\{ \sum_{j=1}^{i-1} n_j - \sum_{j=i+1}^{m} n_j \right\} \tag{6}$$

$$\dot{n}_i = \alpha n_i \left\{ \sum_{j=1}^{i-1} \left(\frac{n_j}{\sum_{k=j}^{m} n_k} \right) - \left(\frac{1}{\sum_{k=i}^{m} n_k} \right) \sum_{j=i+1}^{m} n_j \right\} \tag{7}$$

$$\dot{n}_i = \frac{\alpha}{N} \left\{ \sum_{j=1}^{i-1} n_j^2 - (m-i) n_i^2 \right\} \tag{8}$$

$$\dot{n}_i = \alpha \left\{ \left(\frac{1}{\sum_{k=1}^{i} n_k} \right) \sum_{j=1}^{i-1} n_j^2 - n_i^2 \sum_{j=i+1}^{m} \left(\frac{1}{\sum_{k=1}^{j} n_k} \right) \right\} \tag{9}$$

It is easy to show that model (6) is a generalization for more techniques of the logistic model. Suppose that model (6) describes the actual diffusion process, and aggregate all techniques from 1 to $i-1$ into one category, called the 'old' technology: $n_{old} = \sum_{k=1}^{i-1} n_k$ Aggregate and the rest, from i to m, into the 'new' technology: $n_{new} = \sum_{k=i}^{m} n_k$. Then the diffusion of the new technology is determined by the switching of firms from techniques from the interval 1 to $i-1$ to techniques in the interval i to m. These jumps can be expressed by:

$$\Delta\left(\sum_{j=i}^{m} n_j\right) = \sum_{j=1}^{i-1}\left(n_j \sum_{k=i}^{m} p_{jk}\right) \tag{10}$$

The operator Δ is used to indicate change over time. Substitution of expression (2) yields:

$$\Delta\left(\sum_{j=i}^{m} n_j\right) = \frac{\alpha}{N}\sum_{j=1}^{i-1}\left(n_j \sum_{k=i}^{m} n_k\right) = \frac{\alpha}{N}\left(\sum_{k=i}^{m} n_k\right)\left(\sum_{j=1}^{i-1} n_j\right) = \frac{\alpha}{n_{new}}\left(1-\frac{n_{new}}{N}\right) \tag{11}$$

This is the familiar logistic curve for the 'new' technology. A type of logistic diffusion model similar to equation (10) is analysed by Iwai (1984a and b): not any single technique, but the cumulation of techniques i to m passes through a logistic diffusion process.

The models (6) to (9) produce different diffusion curves. The curves of models (6) and (7) have the well known ogive form, whereas the curves produced by (8) and (9) remind of a more exponential diffusion process, a diffusion process that emanates if it is driven by information that spreads from one central source [see Mahajan and Peterson (1985)]. The four models are illustrated in Figures 2 to 5. At period 0 technique 1 is used by 95 firms and technique 2 by one firm. At period 30, 38, 50 and 75 techniques 3 to 6 respectively are introduced. Parameter α is .1 for all runs.

Diffusion as described by the first model, the push variant, displays a slow take off and a gradual acceleration of the process. Firms are more likely to switch to a new technique once this technique has been introduced by a substantial number of competitors. Diffusion as described by the second model, the pull variant, takes off at high speed, because the attractiveness of the new technique is highest in the first periods after its launch, when still no competitors use it. Model 1b shows a somewhat quicker introduction of new techniques than model 1a. The models 2a and 2b yield very similar graphs, the curves of model 2b being slightly steeper.

All four models have been estimated using data from our case study banking organization. We distinguish four different levels of technology in these banks (see Figure 5 in chapter 4). The first level is characterized by traditional, mechanical, techniques; the second level is characterized by an automated back office, and the third by an automated front office; the fourth level, finally, is characterized by the presence of an automatic teller machine. From our data on automation, we know the changes in the numbers of banks on each level, between 1980 and 1987. The first level has been left behind by every bank in 1984; the third level is first reached in 1981 and the fourth in 1984. This gives us exactly 22 data points that can be used for the estimation of equations (6) to (9), 5 for the first level, 8 for the second, 6 for the third and 3 for the fourth. In equations (6) to (9), the right hand side, except for α, is calculated using data from

Figure 2

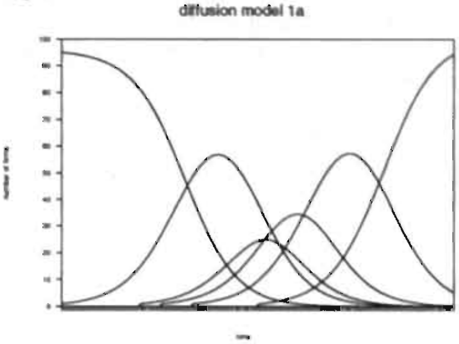


Figure 3

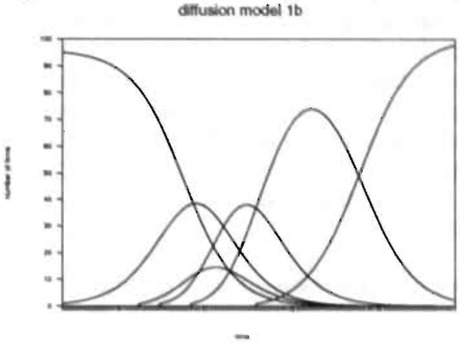


Figure 4

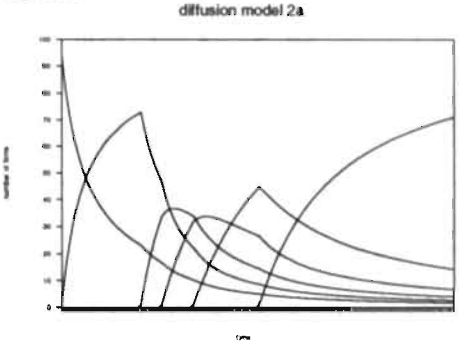
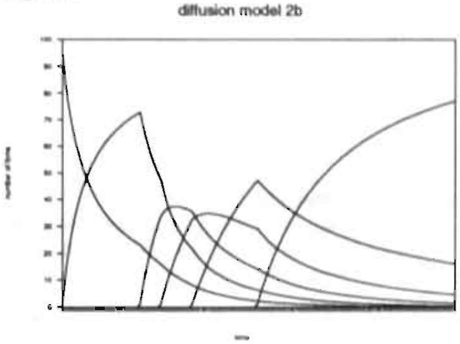


Figure 5



time $t-1$, which serve as explanatory variable for n_{i_t}, the change in the number of banks that use technique i at time t. Then α is estimated using linear regression, assuming that the error term is additive and normally distributed. The outcomes of model (6) resemble very closely the outcomes of model (7). Not unexpectedly, the outcomes of models (8) and (9) are alike. The results are reported in Table 1 and graphed in Figure 6.

Table 1: Results of estimation of equations (6) to (9) in levels

model	α	*adj.*R^2	*D-W stat.*
1a.	.858** (.112)	.738	1.189
1b.	.812** (.116)	.700	.998
2a.	.249** (.045)	.597	.788
2b.	.247** (.043)	.608	.804

standard errors in parentheses;
**: significant at the 5% level; *: significant at the 10% level;
for 22 observations and one regressor, 5% significance points d_l and d_u are 1.12 and 1.31 respectively.

All the dependent variables are highly significant. The models have been estimated with a constant term, but these all proved insignificant. The goodness of fit of all the models is rather high, but the models built on the assumption of a pressure mechanism perform better than the models built on the assumption of a pull mechanism. The Durbin-Watson statistics reported are adjusted for the fact that the sample is pooled out of four sub-samples. These statistics indicate that the residuals are likely to be autocorrelated. Re-estimation of the equations in first differences removes the first order autocorrelation. Calculating first differences costs four data points, one for each technical level, which leaves us with 18 observations. Results can be found in Table 2 and Figure 7.

Table 2: Results of estimations of equations (6) to (9) in first differences

model	α	*adj.*R^2	*D-W stat.*
1a.	.528**	.202	2.160
	(.248)		
1b.	.544**	.204	2.140
	(.253)		
2a.	.156 *	.163	1.581
	(.083)		
2b.	.156 *	.167	1.628
	(.082)		

standard errors in parentheses;
**: significant at the 5% level; *: significant at the 10% level;
for 18 observations and one regressor, 5% significance points d_l and d_u are 1.03 and 1.26 respectively.

The estimates for α are all still significantly positive. A constant term, when included, turned out consistently insignificant. The point estimates for α decrease somewhat and the fit deteriorates, as was to be expected. The graphs show that the models follow the development of the second and the third technological level, on which we have most of the data points, best. The automatic teller machine follows an atypical path of diffusion: in its first year it was adopted by more banks than in the second year, and the upswing came only in the third year.

The conclusion so far must be that these diffusion models do not perform badly in explaining technical progress in the case study bank. If adoptions can be explained to such an extent only from previous adoptions, this gives some support to the hypothesis that banks introduce new techniques, partly because other banks do so. Furthermore, the first type of diffusion model, based on the idea that banks are pushed out of their present technique, and then look for a better alternative, seems to perform slightly better than the second type, which was based on the idea that new techniques attract banks with less efficient techniques indiscriminately.

The diffusion models considered so far are based on different specifications of transitional probabilities p_{ji}. Instead of estimating the diffusion models, in which these probabilities are subsumed, one could also try to estimate the specifications of the probabilities themselves. Suppose that the probability p_{ji_t}, the chance that a firm operating technique j at time $t-1$ will switch to technique i at time t, can be estimated by the *actual* number of firms which operate technique j at time $t-1$ *and* switch to technique i at time t, divided by the total number of firms operating technique j at time $t-1$: $\hat{p}_{ji_t} = \frac{n_{(j_{t-1}\to i_t)}}{n_{j_{t-1}}}$, where a hat indicates that it concerns an estimate.

Figure 6

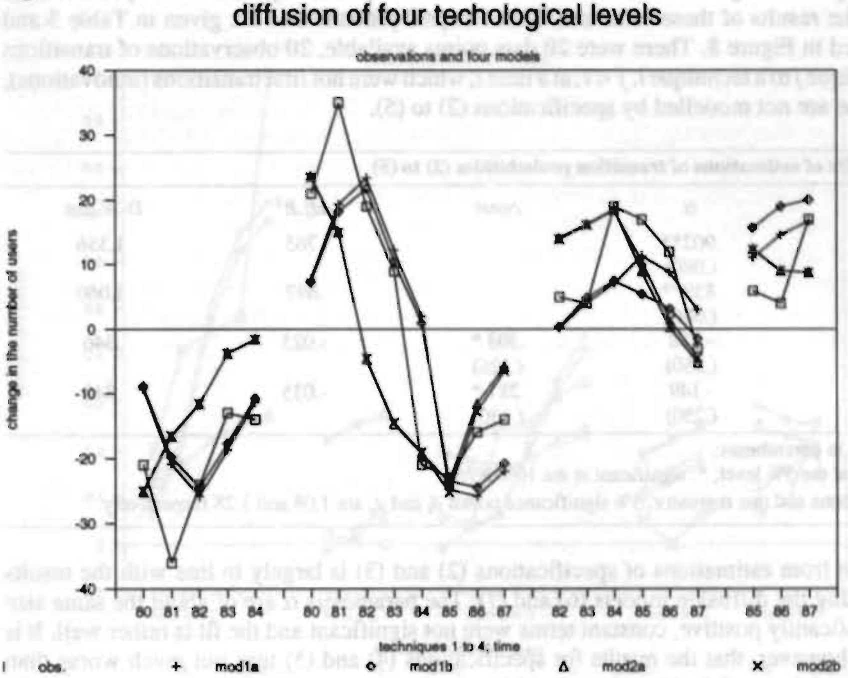
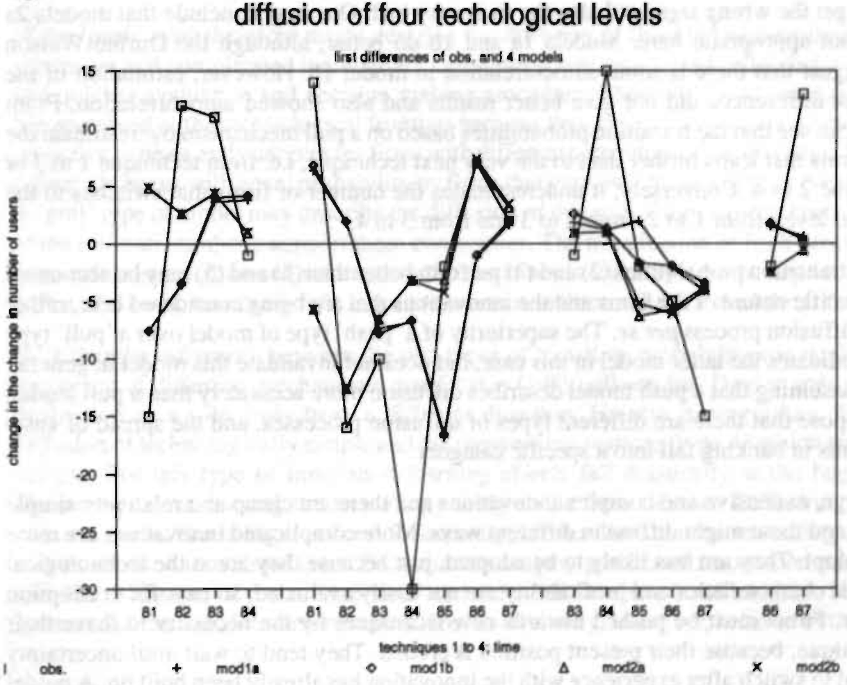


Figure 7



We can see by linear regression how well this estimate is explained by each of the specifications (2) to (5). The results of these estimates of transitional probabilities are given in Table 3 and also displayed in Figure 8. There were 20 data points available, 20 observations of transitions from a technique j to a technique $i, j < i$, at a time t, which were not first transitions (innovations), because these are not modelled by specifications (2) to (5).

Table 3: Results of estimations of transition probabilities (2) to (5)

model	α	*const*	*adj* R^2	*D-W stat.*
1a.	.902**		.765	1.356
	(.080)			
1b.	.839**		.697	1.060
	(.086)			
2a.	-.188	.303 *	-.023	.346
	(.250)	(.126)		
2b.	-.149	.287 *	-.035	.348
	(.250)	(.130)		

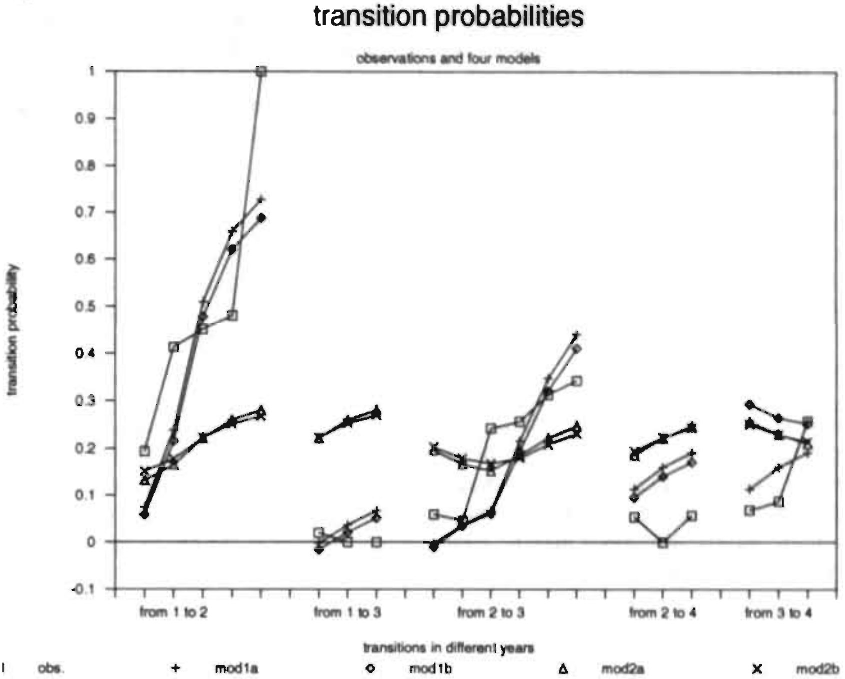
standard errors in parentheses;
**: significant at the 5% level; *: significant at the 10% level;
for 20 observations and one regressor, 5% significance points d_l and d_u are 1.08 and 1.28 respectively.

The evidence from estimations of specifications (2) and (3) is largely in line with the results from estimating the diffusion models (6) and (7). The parameters α are of about the same size and are significantly positive, constant terms were not significant and the fit is rather well. It is remarkable, however, that the results for specifications (4) and (5) turn out much worse than the estimates of the models (8) and (9) would lead one to expect. The parameters α are insignificant and get the wrong sign, and also the fit is very bad. One may conclude that models 2a and 2b are not appropriate here. Models 1a and 1b do better, although the Durbin-Watson statistics suggest that there is some autocorrelation in model 1b. However, estimation of the model in first differences did not give better results and also showed autocorrelation. From Figure 8 we can see that the transition probabilities based on a pull mechanism overestimate the number of firms that leaps further than to the very next technique, i.e. from technique 1 to 3 or from technique 2 to 4. Conversely, it underestimates the number of firms that switches to the next technical level, from 1 to 2, from 2 to 3 and from 3 to 4.

The fact that transition probabilities (2) and (3) perform better than (4) and (5) may be accounted for by the specific nature of the firms and the innovations that are being considered here, rather than by the diffusion process *per se*. The superiority of a 'push' type of model over a 'pull' type of model invalidates the latter model in this case, but does not invalidate this model in general. Rather than assuming that a push model describes diffusion more accurately than a pull model, one may suppose that there are different types of diffusion processes, and the spread of automation systems in banking fall into a specific category.

There are large, expensive and complex innovations and there are cheap and relatively simple innovations, and these might diffuse in different ways. More complicated innovations are more difficult to adopt. They are less likely to be adopted, just because they are at the technological frontier. Their characteristics and profitability are not easily evaluated, so payoffs to adoption are uncertain. Firms must be pushed towards new techniques by the necessity to leave their present technique, because their present position is eroded. They tend to wait until uncertainty decreases and to switch after experience with the innovation has already been built up. A model

Figure 8



of the 'push' type therefore might describe the diffusion of this kind of innovations well. Less expensive and complicated innovations, conversely, will be adopted following a much simpler and quicker evaluation and decision making procedure. They will be adopted as soon as they are perceived at the technological frontier: because they involve a comparatively low risk, it is attractive to be an early adopter for firms with different technological capabilities. Thus adoption is likely to occur with equal probability by firms that vary over the use of their present technique. A 'pull' type of model may describe the diffusion of this sort of innovations best. The outcomes of the estimations above support these conjectures. The introduction of back office automation, front office automation and automatic teller machines is certainly a major step for a local branch office of a bank, involving risk, large investments and retraining of the labour force.

Note that the difference between the two types of transition probabilities is reminiscent of the two types of diffusion mechanisms that Davies (1979) allows for. Davies arrives at the same distinction as we do, only from a different direction. He also distinguishes, first of all, the diffusion of technologically simple and inexpensive innovations (type A) which are usually built off-site. For this type of innovation learning effects fall drastically at the beginning of the diffusion process, leaving the technology fairly stable from then on. Diffusion here follows a concave curve, like in Figures 4 and 5. Secondly, Davies distinguishes technologically complex and expensive innovations (type B), requiring lengthy periods of installation on the adopter's site. Here, because of their 'lumpiness' and the small numbers in which these innovations are produced at the start, learning effects will be slow initially. However, because of the larger scope for improvement of these techniques, the innovations of the last group will overtake the innovations of the first group, both with respect to the rate of diffusion and the ceiling for diffusion. In this case, diffusion follows an S-curve, like in Figures 2 and 3. Examples of items in the first group might be simple supplementary equipment, and examples of items in the second group might be new processes in chemical plants and steel works. The automation systems in banks considered above may also be classified as type B.[2] Davies expresses the difference between diffusion of type A and type B innovations in a probit model framework, by assuming a different type of shifting of the critical size at which a firm will adopt the innovation over time.[3] The actual specification of this shifting mechanism, however, seems somewhat ad hoc. Using these two versions of his probit model, Davies finds empirical support for his classification of innovations and diffusion mechanisms. The approach followed above might constitute an alternative to Davies' approach of dealing with two kinds of innovations. Our models 1 and 2 give similar differences in diffusion curves, but the basis of the distinction between the two types of curves is a different type of mechanism.

The results above can also be compared with results from the marketing literature on diffusion of consumer products. Mahajan and Peterson distinguish between exponential diffusion processes, for items that emanate from one source, and logistic diffusion processes, for innovations that are transferred by contact between adopters. Similar differences in diffusion curves are arrived at, and there too the difference may be related to differences in the characteristics and the complexity of the innovations that diffuse.

6.3 Transition, size and productivity

In section 2 it was assumed that the probability p_{ji} that a firm moves from technique j to technique i depends only on the distribution of firms over techniques, like pictured in Figure 1. Although a large part of the switching of technical levels by our case study banks can be explained by these simple models, it is worthwhile to see whether the estimates of the transition probabilities may be improved by making use of additional information on the banks. A first hypothesis would be that not only the technological position of the firm relative to others in the industry, but also the characteristics of the firm, would determine its probability to switch. If this is the case, then firms that switch from technique j to i on any particular moment t have a different profile than the average of the total group of firms that was using technique j.

Possibly relevant characteristics on which data are available are size and productivity. Larger firms may switch earlier because they can better exploit the innovation and can more easily cope with the risk; less productive firms can gain more by introducing the innovation earlier, though a low productivity may indicate lower technical capabilities. There are no indicators for size and productivity for all 119 banks in the sample used in section 2. For a total of 95 observations though, there are figures for mutations in current accounts Y_1 in 1986 and figures for numbers

[2] Davies distinguishes e.g. the electrical hygrometer in weaving as a type A innovation and automatic track lines in car manufacture, tunnel killns in brick making, and the basic oxygen process and continuous casting in steel production as type B innovations.

[3] Specifically, the critical firm size at which a firm will adopt is $\overline{X} = (\theta_t \varepsilon_i)^{-1/\beta}$, where β is a parameter and ε is assumed to capture firm characteristics and to be lognormally distributed. Then it is assumed that for type A innovations $\theta_t = \alpha t^{\phi}$, and for type B innovations $\theta_t = \alpha e^{\phi t}$.

of current accounts Y_2 in 1986 to serve as size indicator. Also there are for 1986 data of costs of labour which can be used to compute a proxy for productivity. In Table 4, the percentage of firms using a technique *j* at time *t*-1 that switches to a technique *i* at time *t* is indicated.

Table 4: The percentage of firms switching from technique *j* in year *t* to technique *i* in the next.

j	*i*	1979	1980	1981	1982	1983	1984	1985	1986
1	2	22.6	41.5	44.7	50.0	100	.	.	.
1	3	.	.	2.6	.	.	.	.	.
2	3	.	3.3	7.1	5.8	28.0	33.3	29.4	29.2
2	4	.	.	.	.	13.3	3.7	.	8.3
3	4	.	.	.	.	40.0	7.4	9.3	28.6

Legenda: *j*: the old technique; *i*: the new technique.

The numbers in Table 4 differ somewhat from those in Figure 8, because the sample is about 20% smaller. In Table 5, the percentage is indicated that switching firms are larger or more productive than the average firms producing with a certain technique.

The main figures in Table 5 are the ones that come from levels and periods, where a substantial part of the firms jump to another technique. As can be seen in Table 4, these are the jumps from level 1 to 2, in the years 1979 to 1982, the jumps from level 2 to 3, in the years 1983 to 1986, 2 to 4, in 1983, and 3 to 4 in 1983 and 1986. Table 5 shows that switching firms tend to be a little larger than the average of the group they come from: in parts 2 and 3 we see mainly positive numbers, and the few negative numbers that are recorded are comparatively close to zero. This certainly holds for firms that skip one level and go from 1 to 3, or from 2 to 4, but also for firms that go from one level to the next. As far as productivity is concerned, there seems to be no clear tendency: from inspection of Table 5 one cannot infer that less efficient firms switch faster or not.

As shown in chapter 4, the size distribution over banks is rather skewed. To compensate for possible disturbances by outliers, the table has been recalculated using rank numbers for size and productivity instead of values. Results are reported in Table 6. The outcomes confirm the comments made to Table 5. Especially in the earlier years of diffusion, large banks seem to adopt innovations more quickly than small banks.

Table 5: The percentage difference (in various measures) between firms adopting a new technique and the average firm using a certain technique.

	j	*i*	1979	1980	1981	1982	1983	1984	1985	1986
Y_1	1	2	59.2	28.4	.6	24.5	.0	.	.	.
	1	3	.	.	267.3	.	.	.	.	.
	2	3	.	-11.9	92.2	21.8	-1.3	37.2	-6.2	-15.1
	2	4	.	.	.	.	53.6	192.3	.	3.6
	3	4	.	.	.	.	4.8	-3.8	47.1	51.8
Y_2	1	2	52.6	20.2	.1.6	29.1	.0	.	.	.
	1	3	.	.	215.3	.	.	.	.	.
	2	3	.	-1.1	96.9	33.6	-9.2	16.3	7.9	-19.2
	2	4	.	.	.	.	57.4	238.2	.	12.9
	3	4	.	.	.	.	8.0	22.0	19.4	33.2
$\frac{Y_1}{v_0}$	1	2	2.9	-5.3	2.0	5.3	.0	.	.	.
	1	3	.	.	-23.2	.	.	.	.	.
	2	3	.	24.5	-9.8	26.1	-14.8	-4.4	33.9	-2.1
	2	4	.	.	.	.	8.8	-6.4	.	39.0
	3	4	.	.	.	.	-6.5	39.1	-25.5	-2.4
$\frac{Y_2}{v_0}$	1	2	7.3	1.1	1.1	1.6	.0	.	.	.
	1	3	.	.	-10.5	.	.	.	.	.
	2	3	.	10.8	-11.9	14.9	-7.5	12.9	16.4	2.9
	2	4	.	.	.	.	6.2	-19.1	.	27.5
	3	4	.	.	.	.	-9.3	9.7	-8.3	11.3

Legenda: *j*: the old technique; *i*: the new technique;
Y_1: the percentage that firms switching from *j* to *i* are larger than the average firm using technique *j*, where size is measured by mutations in current accounts;
Y_2: like 2., with size measured by number of current accounts;
$\frac{Y_1}{v_0}$: the percentage that firms switching from *j* to *i* are more productive than the average firm using technique *j*, where productivity is measured by mutations in current accounts, divided by labour costs;
$\frac{Y_2}{v_0}$: like 4., with productivity measured as number of current accounts, divided by labour costs.

A second hypothesis to be considered might be the assumption that the weight of a firm for the diffusion process is determined by its size. The choice of production technique of larger firms may influence the transition probabilities more than the choice of small firm. The chance that a firm using technique *j* at time *t* will use technique *i* at time *t*+1 may be larger when the firms already using *i* are large rather than small. If this is the case, then a formulation of equations (2) to (5) in terms of size would yield better results than the estimations in section 2. If *y* stands for output, we get:

Table 6: The percentage difference in rank number (various orderings) between firms adopting a new technique and the average firm using a certain technique.

	j	i	1979	1980	1981	1982	1983	1984	1985	1986
Y_1	1	2	59.8	37.4	10.2	41.5	.0	.	.	.
	1	3	.	.	220.4	.	.	.	.	.
	2	3	.	7.2	30.6	35.5	2.4	33.4	-7.3	-20.1
	2	4	.	.	.	.	48.7	124.2	.	18.1
	3	4	.	.	.	.	7.1	14.0	47.5	47.7
Y_2	1	2	51.4	24.0	6.2	39.9	.0	.	.	.
	1	3	.	.	178.5	.	.	.	.	.
	2	3	.	16.5	38.3	39.8	-8.7	19.3	2.9	-23.6
	2	4	.	.	.	.	53.4	96.6	.	26.0
	3	4	.	.	.	.	4.2	40.9	36.6	37.1
$\frac{Y_1}{y_0}$	1	2	19.7	2.1	4.5	4.8	.0	.	.	.
	1	3	.	.	-67.5	.	.	.	.	.
	2	3	.	33.0	-35.5	45.3	-12.6	6.5	24.2	-15.0
	2	4	.	.	.	.	20.2	-43.8	.	65.8
	3	4	.	.	.	.	-6.4	14.9-	27.0	5.9
$\frac{Y_2}{y_0}$	1	2	5.6	-8.8	5.3	-2.9	.0	.	.	.
	1	3	.	.	-39.4	.	.	.	.	.
	2	3	.	34.1	-9.0	30.6	-26.0	-10.6	16.3	-3.7
	2	4	.	.	.	.	22.0	-28.1	.	25.1
	3	4	.	.	.	.	-10.5	85.8-	41.7	-7.1

Legenda: see Table 5.

$$p_{ji} = \alpha \frac{y_i}{Y} \tag{12}$$

$$p_{ji} = \alpha \frac{y_i}{\sum_{k=j}^{m} y_k} \tag{13}$$

$$p_{ji} = \alpha \frac{y_j}{Y} \tag{14}$$

$$p_{ji} = \alpha \frac{y_j}{\sum_{k=1}^{i} y_k} \tag{15}$$

Here $Y = \sum_{k=1}^{m} y_k$ is total output. Equations (12) to (15) have been estimated in logarithmic form, using data from 95 banks, where size has been approximated again by mutations in current accounts Y_1 and numbers of current accounts Y_2. Table 7 gives a summary of the outcomes.

Table 7: Results of estimations of transition probabilities (12) to (15) for 2 measures of size.

	mod.	α	*const*	*adj.*R^2	*D-W stat.*
n	1a.	.872**		.723	1.184
		(.083)			
	1b.	.827**		.681	1.026
		(.086)			
	2a.	-.167	.297**	-.031	.340
		(.252)	(.123)		
	2b.	-.128	.284**	-.041	.339
		(.255)	(.128)		
Y_1	1a.	.713**		.659	919
		(.080)			
	1b.	.679**		.614	.819
		(.083)			
	2a.	-.202	.297**	-.009	.343
		(.221))	(.095)		
	2b.	-.174	.290**	-.021	.340
		(.225)	(.099)		
Y_2	1a.	.724**		.650	.874
		(.083)			
	1b.	.682**		.588	.763
		(.086)			
	2a.	-.204	.299**	-.011	.342
		(.229)	(.099)		
	2b.	-.170	.290**	-.025	.339
		(.233)	(.104)		

standard errors in parentheses;
n : number of firms (compare Table 3);
Y_1 : size measured as mutations in accounts;
Y_2 : size measured as number of current accounts;
**: significant at the 5% level; *: significant at the 10% level;
for 20 observations and one regressor, 5% significance points d_l and d_u are 1.08 and 1.28 respectively.

The first part of Table 7 is the equivalent of Table 3, using only data for 95 banks. The results in the second and the third part are rather similar to the ones in the first part and in Table 3. Models 2a and 2b also perform as badly as before when taking account of firm size; models 1a and 1b perform reasonably well, although first order autocorrelation may be present. The results indicate that it is hard to improve on the explanatory power of the generalized logistic model, equation (6), for the diffusion of new technologies within our case study bank. Apparently the size of a bank does not contribute much to the weight it has in the diffusion process. It may be concluded from this section that the data suggest that the probability that a bank switches over to a new technique only depends on its own size and the distribution of all banks over techniques. Larger banks, on average, seem to switch faster to a new technology than smaller banks, but the size of a large bank does not seem to be of influence on the switching of other banks.

6.4 Conclusion

The decisions of firms to introduce a new production technology are likely to be based on a variety of considerations. Firms will attempt to make an estimate of costs and benefits of adopting an innovation. The benefits can be growth of production capacity and market share and an increase in efficiency. The costs are an investment in technology, but also in training of labour and in evaluation and decision making itself. As innovations get more complex and their impact on the activities of the firm get more profound, it gets more difficult to estimate the benefits of adoption and the costs of evaluation and decision making start to rise as one tries to reduce uncertainty. Under these conditions, it is likely that firms will try to learn from each other. The adoption policy of competitors conveys information on their estimates of the balance of costs and benefits of an innovation. Moreover, generally it is not efficiency in absolute terms that counts for survival, but one's efficiency in comparison to competitors (at least in case demand is not very price elastic). Therefore, firms may adopt merely to reduce the risk of falling behind, in the hope that adoption will pay off at some later date. In this way, the main motor of diffusion becomes diffusion itself.

In this chapter, models have been built on the assumption that bandwagon effects are the main determinants of diffusion. Two types of diffusion mechanism were suggested, one a 'push' and the other a 'pull' mechanism. The push mechanism expresses that firms start to innovate, when they begin to loose ground with their present routine and are pushed to improve efficiency. The pull mechanism expresses that new techniques conquer the market by attracting potential users from all other techniques indiscriminately, because their superiority to perform a specific function can be easily evaluated and the risks of adopting are low. Both mechanisms were translated into two models of the probability of transition from an inferior to a superior technique. The resulting four formulations of transition probabilities were then used to construct four diffusion models. These diffusion models were estimated. The difference between the results obtained from models built on the assumption of a push mechanism were roughly equivalent, maybe slightly better, than the results obtained from models built on a pull mechanism. The data set at our disposal, however, does not only contain information on the total number of users per technique per year, but also more detailed information on what firm switches at what time from which old technique to which new. This allowed for a direct estimation of the transition probabilities themselves and a test of the four models for these probabilities. Here the push models performed considerably better than the pull models. This is not surprising, considering the character of the innovations which are represented in the data, which can be described as complex, expensive and risky. Nevertheless, it is remarkable that it is so difficult to discriminate between the two types of models, when using only aggregate data on the use of techniques, whereas the difference between the two is so manifest when one can take account of actual transitions between techniques.

Taking advantage of the range of data available, an attempt was made to improve the estimation results by making use of information on the size of banks in the model. The size of a bank, however, does not seem to influence its weight for the diffusion process. On the other hand, it was shown that on average larger firms tend to lead slightly in the adoption of innovations.

A comparison of the empirical results of this chapter with those of the foregoing chapters suggests a conclusion on the relative importance of the two types of considerations mentioned for the adoption of automation equipment in our sample of banks. A firm's decision to adopt an innovation may be based partly on a firm's assessment of its production costs relative to the benefits, and partly on bandwagon-effects, on perceived decisions of competitors. The analysis in chapter 5 was based on the assumption that firms invest in new technology on the basis of expected costs and benefits. The analysis in this chapter takes the hypothesis of bandwagon effects as its basis. The data analysis in chapter 4 and the estimation results in chapters 5 and 6 suggest that, for the sample of banks for which data were available, adoption of a technology in reaction to competitors adoptions was much more typical than adoption on the basis of a cost benefit analysis only. Whereas little relationship between productivity, costs and investments in new technology could be found, fairly regular diffusion patterns of innovations turn up in the data. This relative importance of bandwagon effects for the adoption of new technologies by this sample of banks can be explained various factors. First of all, because the technologies that banks are deciding on are expensive and complex, and have pervasive consequences for the production processes, adoption behaviour of other firms in the industry may be an important source of information. Secondly, because the banks belong to one large organization of banks, there may be some degree of coordination and there may be growing network externalities to adoption.

7. Summary and concluding remarks

7.1 Introduction

Economics is closely linked to technology: technology is the way we transform the scarce resources with which we attempt to fulfil our unlimited needs. Though in the short term technology may be considered a datum, it is subject to change in the long term. Technological change both causes changes in the constraints of the economic system and is itself induced by economic behaviour. To deal with long term economic development, economic analysis therefore needs the conceptual and mathematical tools to deal with technological change. A number of useful conceptual tools can be found in evolutionary economics: the concepts of technological trajectories and paradigms, path dependence, dynamic adjustment, complexity and uncertainty, routines and bounded rationality (see Table 1 in chapter 1). These concepts, however, which refer to intuitively important aspects of technological development, are not well integrated into mainstream (micro-) economic thought and modelling practice. This study is an attempt at a contribution to such an integration.

7.2 Subject outline

Firms invest in new production techniques, and thereby alter their production capacity, their demand for inputs and their profitability. This thesis tries to contribute to our comprehension of this process. The central issue is to understand how firms decide on investments in new production technology. In particular, it is important to specify the *choice opportunities* of the firm, its *objectives* and its *methods*. The analysis has concentrated on opportunities and methods, accepting profit maximization as the firm's main objective. The choice opportunities open to the firm are determined by its history; thus development is path dependent. The firm's method of handling information and taking decisions is tied to the limitations of its cognitive capacity; thus perception is selective and rationality is bounded.

The study aims at three goals. The first is to contribute to the conceptual analysis of some aspects of technological progress and of its role in economic development. The second goal is to formalize these theoretical reflections in models and to analyse the theoretical postulates with the help of these models. The third purpose is to use the models as a tool in empirical verification of the theoretical insights. This chapter first summarizes the main aspects of the thesis. It then concludes with some general remarks and assessments.

7.3 Summary

Chapter 1 introduces the subject of the thesis, mentions the various relevant theoretical perspectives and gives a brief outline of the things to follow.

Chapter 2 provides the theoretical background of the study. It starts out with a general depiction of economic behaviour. Economic activities of production, exchange and consumption require cognitive activities to exert control. Depending on the character of the economic activity, these cognitive activities range from very simple (buying an ice-cream) to very complex (working out a corporate strategy). As these cognitive activities get more complex, they impose binding constraints on economic behaviour. The cognitive resources of decision makers are limited, and therefore choices have to be made on where to employ them. The economic problem of choice under conditions of scarcity is extended from the real world to the cognitive realm: a preliminary choice problem in the 'control sphere' becomes which choices in the 'real sphere' to produce, which knowledge to generate, which information to evaluate, which skills to develop. When firms try to improve their production technologies, cognitive constraints are likely to become binding and to co-determine the direction and speed of progress. It is important to note the differences and similarities between the economic problem of the 'real sphere' and the economic problem of the 'control sphere', for these indicate to what extent the same tools of analysis may be applied to both spheres of economic activity. Production of goods and services in the real sphere is matched by production of information and decisions in the control sphere. Both types of production require investments of scarce resources and foregoing consumption opportunities. An important resource in both spheres is time: investments in both spheres have a gestation time. The 'products' in both spheres, however, differ in character, mainly because cognitive products are less appropriable.

After this general sketch, chapter 2 turns to the field that further analysis will concentrate on: the introduction of new technology. Among other topics, the relationship between innovation adoptions and the diffusion process is explored, and the question emerges where the empirical regularity of the diffusion curve originates, at the micro level or at the industry level. Either a regular distribution over firms of the characteristics that determine adoption decisions may explain a regular diffusion curve, or the regularity of the diffusion curve may be predominantly determined by the mutual dependence of adoptions, in the sense that every adoption provokes the next, although one cannot determine in advance the sequence in which firms will adopt.

Chapter 3 elaborates on the verbal description of economic behaviour in chapter 2 and tries to capture the main thoughts in a model of firm decision making. The firm is assumed to produce 'real sphere' output by means of a constant technique and in constant volumes. In the control sphere, the firm produces decisions, information, new technology or whatever is necessary, in order to increase its future profitability. Future profitability depends on production capacity and on the future productivity of those inputs that have not yet been bought, but will be necessary in production: the productivity of variable production inputs. Together, output, growth in production capacity and variable factor productivity are the products of the firm: on a 'lower' level the firm produces output, on a 'higher' level it produces itself, its own production possibilities. These higher level production processes are described by two production functions. Their input is aggregate gross investment, which is invested both in the control sphere (R&D, decision making, training) and in the real sphere (production equipment). Because at any moment in time, resources for investment are scarce, have a declining marginal productivity and have two alternative employments, profit maximizing firms are assumed to choose a rate of investment and a ratio of producing capacity growth and producing productivity growth.

The basis of the model is kept very simple. There is no technical depreciation, no market for second hand capital goods, no vintage structure of the capital stock, no financing restriction, no change in the price of investment goods. The model of the production of final output is rigid and as simple as possible. Any change at the lower production level of the firm originates from activity at the higher production level. At this higher level of production, the firm is engaged in a prolonged process of changing its lower level production constraints, of adjustment to changing opportunities and market conditions. At the lower level, markets clear and the economy is permanently in momentary equilibrium; at the higher level there is a process of on-going adjustment. A next issue is to introduce technological change in this model. It is argued that the determining characteristic of technological progress is its cumulative character, which can be associated with the concept of learning and with dynamic economies of scale. This cumulativeness implies that the returns to current investment in technological change depend on the extent of investment in both the future and the past. The decision problem of the firm is therefore necessarily a dynamic planning problem: the planning of a time path for investments.

The most important elements of the model can be summarized as follows. The firm maximizes the present value of its future income, given market, technological and cognitive constraints. The latter constraints can be characterized by four assumptions. It is assumed, first of all, that without investments the firm cannot change production capacity nor production technique; secondly, that investments allow the firm to expand capacity and to increase variable factor productivity; thirdly, that cognitive constraints restrict the firm in changing scale of operations and production technique; fourthly, that if the position of the firm on its technological trajectory allows for technological change, then investments have a cumulative effect.

This model was used to examine the processes of investment planning and of competition, both analytically and by simulating. The focus was on the course of the development path of firms and of an industry at large. It was found that firms, given perfectly competitive input and output markets and technological opportunities, tend to plan investments *growing* without bound over time and moving gradually from factor productivity growth toward capacity expansion. However, when markets are not perfectly competitive (expressed by finite price elasticities), this conclusion no longer holds: investments will then continue to be directed at improving efficiency and are likely to *decrease* in the long term. In a particular case, if case price elasticities of inputs and output are constant, a warranted steady state rate of growth of output can be derived. However, there is not only no mechanism that ensures convergence of planned growth rates to the warranted rate, but it is also shown that the warranted rate itself is unstable.

One variable on which investment plans of firms depend is the expected behaviour of competitors on input and output markets. It was shown that firms of different size are likely to react differently if they expect competing supply to increase faster in the future. Whereas it is optimal for small firms to decrease spending on expansion of capacity, it is rational for large firms to plan a retaliatory increase in capacity. The threshold size above which firms retaliate as they expect more competition depends on their market share and on the price elasticity of output demand.

The model has been refined to analyse competition under different conditions on markets for inputs and outputs. Development of firms and industries in this model depends crucially on starting conditions: future development cannot be seen separate from historical development. The model only describes rational planning, given a starting condition.[1] As there can be many different starting conditions, the model can accommodate many different types of development, in which different firms survive side by side for lengthy periods of time with different market shares, different growth rates, different levels of efficiency and different speeds of technological progress. Moreover, variables like growth rates, market shares and profits can diverge and converge at different stages of industry development. An example of this was elaborated with the help of a computer simulation model, simulating the competitive process of a small number of different firms conquering a new market.

It was shown that, without entry, if the demand curve for output is downward sloping, investment halts at some time, despite positive profits. Over time, the industry develops in accordance with elements from the *product life-cycle theory*: output prices fall gradually, investment first rises and then decreases, the industry moves from an expansionary stage to a stage of rationalization and consolidation, and the largest profits are made in the periods of sharpest market growth. Thus the product life cycle is generated in this model by the process of technological progress and microeconomic competitive behaviour. The marketing strategy known as 'skimming' appears as an optimal pricing strategy.

The relationship between *firm size, market structure and the speed of technological progress* turned out to be intricate in this type of model, because it is co-determined by the price movements and price elasticities of inputs and output. Firm size (absolute size) appears to vary positively with investment in technology, because larger firms reap scale benefits from the introduction of new technology. At the same time, market share (relative size) varies negatively with investment in technology, because as firms have a higher share of the market with a given demand elasticity of output, they internalize a negative price change of output to a larger extent. The influence of firm size and market structure on the speed of progress varies over the life-cycle of the market or industry. A fall in the output price reduces the power of the firm size factor and a rise in the elasticity of demand increases the influence of the market share argument. Firm size dominates the determination of the speed of technical progress in a developing market and market structure dominates in a mature market.

The interplay of the firm size and the market share effect on investment in capacity expansion also explain the appearance of a *sigmoid diffusion curve* of output. As a market opens up, firms grow and thereby increasingly create the opportunity to exploit economies of scale. This induces them to expand at an increasing rate. As the market saturates and the price elasticity of output demand rises, the output price tends to fall rapidly as output expands, putting profits increasingly under pressure. This induces firms with a large market share to decrease the rate at which they increase output supply.

[1] One may wonder what explains the starting condition, and whether the starting condition is likely to fulfil certain characteristics. Note that in a history dependent world, the conditions at the start are path dependent, determined by the historical process to date. Therefore there is no reason to assume that starting conditions should exhibit some general pattern.

Chapter 4 is the bridge from the theoretical reflections to the empirical evidence. To find empirical support for the proposed modelling approach, data were used from the Dutch banking industry. The banking industry is an important part of the services sector, in which a number of recent trends, which are prominent in large parts of the services sectors of western economies, are particularly visible. Services industries, on the one hand, gain in importance relative to manufacturing industries, e.g. in terms of share of gross national product and employment, and, on the other hand, start to adopt more and more features that used to be characteristic for manufacturing industries. Services enterprises get more capital intensive, standardize their products, compete by diversification and market segmentation, expand and merge to be able to cover large markets and to reap economies of scale and scope. An important cause of this trend is the development of computer and telecommunications technologies. The banking industry, a large-scale consumer of information technology, exemplifies this trend very well. The application of information technology in banking has lead to the development of a large range of new products for the retail and the wholesale market, to a large expansion of output and a dramatic decrease in costs.

A set of data from local branch offices of a Dutch banking organization was available to be used for testing the models. In chapter 4 these data are first analyzed inductively, showing amongst other things sigmoid diffusion curves for automation equipment and indications of scale economies in production. To test the models of chapter 3, variables have been selected from this data set to approximate output, variable inputs, investment and technological level. The main investments in technological progress in banking organizations at the local level over the period 1979 to 1987 were in automatic data processing. At first local banks invested in automation systems for back office operations, later in automation systems at the front office, and after that in automatic teller machines and related network systems. The most voluminous product of these local banks, and the one most affected by the introduction of information technology, is the administration of accounts, including the transfer of money from one account to the another. Although the production process is relatively capital intensive, the main inputs into production of this output are still different types of labour: counter work, data input work, administration, management. None of these labour inputs are fully variable, but probably some of them are more variable in the short term than others. Consequently, data on administration of payments were chosen as proxies for output, different types of labour data as variable input proxies, cumulated investments in automation equipment as investment proxy and availability of three different automation systems as technology indicator.

Chapter 5 is devoted to tests of the model of chapter 3. Because the available data are mainly cross-section data and cover no more than a couple of years, the multi-period model had to be simplified to a two-period model. The scope of the tests of the model is therefore limited to certain aspects. Not so much the assumed mechanism of investment itself as the extent to which the mechanism is the same for different banks is tested. The first question dealt with is, whether there is evidence that investments have a cumulative effect, whether they lead to technical change, rather than to pure expansion of the capital stock. The second question is, in how far technological opportunities for different firms in the same industry at different moments can be represented by a single stable function, to what extent different firms face the same technological constraints.

The outcomes of the estimations have to be considered with care, because the reliability of the data as proxies for the concepts of output, variable inputs and investment is not unconditional and because the model is a very stylized representation of a banking firm. Nevertheless, some positive results deserve to be reported. Estimations lend support to the hypothesis that all banks face sharply declining marginal returns to investment in automation at any period in time. In spite of that, large banks invest more than small banks. Parameter estimates indicate that this may be explained by the fact that large banks can take advantage of scale economies in investment. They reap more benefits in absolute terms from any level of investment, and can therefore spend more on investment before marginal returns equal marginal costs. This lends tentative support to the hypothesis that investments have a cumulative effect, which may be because investments change over-all production technology. The relationship between investment and demand for variable inputs was less consistent, but does not exclude a cumulative effect.

Concerning the question to what degree different banks face the same constraints and follow the same technological track, only preliminary answers could be found. Although all firms take the same qualitative steps in technological development, by first introducing back office automation, then front office automation, and after that linking up to the network of automatic teller machines, they do not seem to follow this path with the same speed. Also firms do not move in a particular order. Considering quantitative variables, firms do not invest in the same direction, nor do they invest similar amounts in automation equipment; rather they invest at roughly the same rate. Although the constraints to their development, in terms of the quantitative variables output, inputs and investment, could be estimated, under the assumption that the main parameters were the same for all firms, a number of analytically derived hypotheses concerning the direction of investment could not be supported by evidence.

Chapter 6, finally, starts out from the observation that no relationship between costs, benefits and the introduction of specific technological system was found in chapter 5, but that the spread of these systems through this group of banks nevertheless showed a regular diffusion pattern. We therefore reverted to a theme introduced in chapter 2, to the question where the regular shape of the diffusion curve has its roots, at the micro level or at the aggregate level. If at the micro level the tie between capacity and productivity on one side and investment behaviour and adoptions of innovations on the other is so ambiguous, a closer look at possible mechanisms at the aggregate level is warranted. Aggregate level mechanisms may be dominant in determining the shape of technological progress in an industry, when bandwagon effects are important, or when individual behaviour is erratic, but the market operates as a selection mechanism once firms have set out on an investment course. The mechanisms explored in chapter 6 assume that the relative profitability of a technique, and the relative frequency of its use, determines the probability that a firm using a less profitable technique will adopt it and a firm currently using it will leave it for a more profitable alternative. These probabilities may be used to formulate a diffusion model. A distinction is made between a 'push' and a 'pull' mechanism, depending on whether a firm is supposed to decide to switch because it is driven out of its old technique, or because it is attracted to a new. The two different mechanisms result in different types of diffusion curves, the push-curves being sigmoid and the pull-curves positively sloped and concave. It may be expected that the push type of mechanism will be operative in the diffusion of complicated expensive innovations and the pull type in the diffusion of relatively simple and cheap innovations. The two shapes of the curves are reminiscent of the diffusion curves of Davies' type B and type A innovations, but the supposed mechanisms are different.

The 'push' and 'pull' diffusion models have been estimated using the data of innovation adoptions from the sample of banks. Although a test of two diffusion models based on a push and a pull mechanism respectively did not give radically different results, a direct test of the transition probabilities according to the two mechanisms showed large differences. The latter indicated that a push mechanism is more likely to be important in these banks than a pull mechanism, an outcome which could be expected from the character of the innovations under consideration. A comparison of the empirical outcomes of chapter 5 and 6 suggests that, next to considerations of capacity growth and productivity increase, aggregate mechanisms like bandwagon and spill-over effects were important determinants for the introduction of information technology in banks over the last decade.

7.4 Concluding remarks

Key assumptions in this thesis are:

1. Development is history dependent; investment decisions concern incremental changes.
2. Cognitive constraints are an important determinant for the process of economic and technological development.
3. Technological change is driven by economic motives; technological opportunities are given by the nature of trajectories.

The *main conclusions* from the study may be briefly summarized under six headings.

First, an explanation for the product life-cycle and the sigmoid diffusion curve may be found in the process of competition as described by the firm investment planning model (chapter 3).

Secondly, whether a small number of competitors generate faster technological progress than a large number of competitors in a market (cf. the so called Schumpeterian hypotheses) depends, on the one hand, *positively* on their possibility to realize scale economies in technological progress and, on the other hand, *negatively* on the extent to which they internalize price effects. The first factor depends on their *size*, the second on their *market share*. Moreover, the relationship between the speed of technological progress and market structure may differ for investments in capacity expanding and variable input saving technology. Finally, the relative strength of the size and the market share effect change over the course of the industry life-cycle, as prices fall and the elasticity of demand for output goes up (chapter 3).

Thirdly, there is no steady state growth path for any single market or industry producing a homogeneous output. If output markets are infinitely elastic, or the number of firms in the market approaches infinity, investment and the rate of output growth increase without limit; if the demand curve is downward sloping (not asymptotically horizontal), investment goes to zero; if the elasticity of demand is constant and above unity, a warranted rate of steady state growth may be computed, which is unstable, however (chapter 3).

Fourthly, estimation results indicate that in the sample of banks we considered there are sharply decreasing marginal returns to investment. They also point at important scale effects to investment: the same amount of investment generates more returns (in absolute terms) to a larger firm. These scale effects may reflect technological change (chapter 5).

Fifthly, in our sample of banks, the measurable relationship between costs, benefits and the adoption of specific technical systems appears weak (chapter 5).

Sixthly, bandwagon and spill-over effects appear to contribute importantly to the explanation of the adoption of technical systems by the banks in the sample. The diffusion patterns fit the model specification for more substantial, high-risk innovations better than the model specification for incremental, low-risk innovations. This accords positively with intuition, since the technical systems under consideration are relatively complex and influence banking operations significantly (chapter 6).

The thesis leaves many matters pertaining to its declared topic open, undecided or up in the air. These may become subjects for further research. A short list of the more theoretical issues for future exploration might include:

1. The firm model of chapter 3 can be evaluated under different expectations regimes, e.g. under the assumption of rational expectations and strategic behaviour.

2. The relationship between market structure and technological change can be evaluated, given not only an indirect external effect from technological change through prices, but also the direct external effect of technology spill-overs.

3. In the stylized world of chapter 3, the firm is connected to its environment through the market for output, the market for variable inputs and the credit market. The interest rate has been assumed constant and thus the supply of credit infinitely interest elastic; capital goods were supposed to be supplied at constant prices; the demand curve for output and the supply curve for variable inputs were assumed exogenously determined and fixed over time. To examine competition and development in a more dynamic environment, planning models for suppliers of variable inputs or labour and for consumers of output could be specified to supplement the firm model, thereby endogenizing the position of the input supply and output demand curves. Similarly the supply of credit and investment goods can be modelled, thus endogenizing capital embodied technological change. These models can be based on the same principles as the firm model: path dependency and limited cognitive capacities.

4. The relationship between microeconomic adoption behaviour and aggregate patterns of diffusion, in cases where adoption seems erratic, but diffusion fairly regular, requires further analysis. The link between microeconomic and aggregate behaviour remains a dismal issue.

On the empirical side several matters comes to mind:

1. An important issue for the empirical usefulness of the model is a certain stability of the functions that express technological opportunities, $g(\beta)$ and $h(M)$. The model is a better tool, when these functions are stable across firms and over time. In the banking industry no evidence for this stability turned up, but banking may not be a typical case.

2. In assessing the relationship between market structure and technological progress empirically, it may be useful to differentiate between technical progress directed at capacity expansion and technical progress directed at variable input saving. Based on the theoretical results above, one might expect that the former goes faster in competitive markets and the latter in more concentrated markets. Moreover, it may be fruitful to take account of the life-cycle stage of the market. Technical progress leading to

expansion would be expected to dominate input saving progress in markets in early stages of the life-cycle, whereas the opposite would hold for markets at a later stage of the life-cycle.

3. A possible use for the firm planning model and similar models for other functions in the economic system would be to build simulation systems of micro models, appropriate for evaluating effects of economic policy. The parameters of the ingredients of the systems, the micro-models, can be estimated with standard procedures. The micro-models can then be allowed to interact through markets, much like was done in the simulation procedures in chapter 3, to assess characteristics of future economic development and consequences of policy.

A source of inspiration for the study of the relationship between technological development and economic dynamics is the urge to escape from the spell of 'technological determinism' and to develop effective policies to guide long term economic development in favourable directions. In this sense, the problem of the links between technological and economic development is an appearance in the economic realm of the problem of free choice, the question if and to what degree man can determine his own fate. A first step to a solution of the problem of free choice is the development of an understanding of opportunities and constraints. May this thesis be a contribution to the construction of tools to explore the technological constraints to choice and their determinants.

REFERENCES

Abernathy, W.J., *The Productivity Dilemma: Roadblock to Innovation in the Automobile Industry*, The John Hopkins University Press, Baltimore and London, 1978.

Ahmad, S., On the Theory of Induced Innovation, *The Economic Journal*, vol. 76, 1966, pp. 344-357.

Arthur, W.B., Competing Technologies, Increasing Returns, and Lock-in by Historical Events, *The Economic Journal*, vol. 99, 1989, pp. 116-131.

Arthur, W.B., Competing Technologies: An Overview, in: Dosi, G. *et. al.* (eds.), *Technical Change and Economic Theory*, Pinter Publishers, London and New York, 1988.

Baldwin, W.L. and J.T. Scott, *Market Structure and Technological Change*, Harwood Academic Publishers, Chur, 1987.

Bank for International Settlements, *Payment Systems in Eleven Developed Countries*, Bazel, 1985.

Bass, F.M., The Relationship between Diffusion Rates, Experience Curves and Demand Elasticities for Consumer Durable Technological Innovations, *Journal of Business*, vol. 53, no. 3, pt. 2, 1980,

Baumol, W.J., J.C. Panzar and R.D. Willig, *Contestable Markets and the Theory of Industry Structure*, Harcourt Brace Jovanovich, inc., New York, 1982.

Binswanger, H.P. and V.W. Ruttan, *Induced Innovation, Technology, Institutions and Development*, The John Hopkins University Press, Baltimore and London, 1978.

Boyer, R., Technical Change and the Theory of 'Régulation', in: Dosi, G. *et. al.* (eds.), *Technical Change and Economic Theory*, Pinter Publishers, 1988.

Central Bureau of Statistics, The Netherlands, *Statistical Yearbooks*, various issues, Voorburg, Heerlen.

Central Planning Bureau, The Netherlands, *Central Economic Plan*, various issues, The Hague.

Cohen, W.M. and R.C. Levin, (1989), Empirical Studies of Innovation and Market Structure, in:R. Schmalensee and R.D. Willig (eds.), *Handbook of Industrial Organization*, North Holland, Amsterdam

Cohen, W.M. and D.A. Levinthal, Innovation and Learning: the Two Faces of R&D, *The Economic Journal*, vol. 99, 1989, pp. 569-596.

Conlisk, J., A Neoclassical Growth Model with Endogenously Positioned Technical Change Frontier, *The Economic Journal*, vol. 79, 1969, pp. 348-362.

Coombs, R., P. Saviotti and V. Walsh, *Economics and Technological Change*, Macmillan, 1987.

Dandrakis, E.M. and E.S. Phelps, A Model of Induced Invention, Growth and Distribution, *The Economic Journal*, December 1966, pp. 823-839.

Dasgupta, P. and P. Stoneman, *Economic Policy and Technological Performance*, Cambridge University Press, Cambridge, 1987.

David, P.A., A Contribution to the Theory of Diffusion, Stanford Centre for Research in Economic Growth, *Memorandum* no. 71, 1969.

David, P.A., *Technical Choice, Innovation and Economic Growth*, Cambridge University Press, 1975.

David, P.A., Clio and the Economics of QWERTY, *American Economic Review*, vol. 75 no. 2, 1985, pp.332-337.

David, P.A., Technology Diffusion, Public Policy, and Industrial Competitiveness, in: Landau and N. Rosenberg (eds.), *The Positive Sum Strategy: Harnessing Technology for Economic Growth*, National Academy Press, Washington, 1986, pp.373-391.

David, P.A., *The Reaper and the Robot: The Diffusion of Microelectronics-Based Process Innovations in Historical Perspective*, Center for Economic Policy Research, Stanford, working paper no. 1, 1984.

David, P.A. and T.E. Olsen, *Equilibrium Dynamics of Diffusion when Incremental Technological Innovations are Foreseen*, Center for Economic Policy Research, Stanford, publication 67, 1986.

Davies, S., *The Diffusion of Process Innovations*, Cambridge University Press, Cambridge, 1979.

Day, R.H., Disequilibrium Economic Dynamics, a Post-Schumpeterian Contribution, *Journal of Economic Behaviour and Organization*, vol. 5, 1984.

Day, R.H., The General Theory of Disequilibrium Economics and of Disequilibrium and of Economic Evolution, in: D. Batten, J. Casti and B. Johansson (eds.), *Economic Evolution and Structural Adjustment*, Springer-Verlag, Berlin, 1987.

Diederen, P.J.M., R.P.M. Kemp, J.Muysken and G.R. de Wit, Diffusion of Process Technology in Dutch Banking, in: Nakicenovic, N. and A. Grübler (eds.), *Diffusion of Technologies and Social Behaviour*, IIASA, Springer Verlag, Berlin, 1991; also in: *Technological Forecasting and Social Change*, vol. 39, 1991, pp. 201-219.

Diederen, P.J.M., R.P.M. Kemp, J.Muysken and G.R. de Wit, Diffusion of Information Technology in Banking: The Netherlands as an Illustrative Case, in: Freeman, C. and L. Soete (eds.), *New Explorations in the Economics of Technological Change*, Pinter Publishers, London and New York, 1990.

Dosi, G., C. Freeman, R. Nelson, G. Silverberg and L.Soete (eds.), *Technical Change and Economic Theory*, Pinter Publishers, London and New York, 1988.

Dosi, G., The Nature of the Innovative Process, in: Dosi, G. *et. al.* (eds.) *Technical Change and Economic Theory*, Pinter Publishers, London and New York, 1988.

Dosi, G. Technological Paradigms and Technological Trajectories, in: C. Freeman (ed.), *Long Waves in the World Economy*, 1983.

Dosi, G., The Research on Innovation Diffusion: an Assesment, in: N. Nakicenovic and A. Grübler (eds.), *Diffusion of Technologies and Social Behavior*, IIASA, Springer-Verlag, Berlin, 1991.

Elster, J. *Explaining Technical Change*, Cambridge University Press, 1983.

Freeman, C. *The Economics of Industrial Innovation*, Frances Pinter, London, 1982.

Freeman, Ch., Diffusion: The Spread of New Technology to Firms, Sectors, and Nations, in: A. Heertje (ed.), *Innovation, Technology and Finance*, Basil Blackwell, Oxford ,1988.

Freeman, C. and C. Perez, Structural Crises of Adjustment: Business Cycles and Investment Behaviour, in: Dosi, G. *et. al.* (eds.) *Technical Change and Economic Theory*, Pinter Publishers, London and New York, 1988.

Freeman, C., J. Clark and L. Soete, *Unemployment and Technical Innovation*, Frances Pinter, London, 1982.

Freeman, C. and L. Soete (eds.), *Technical Change and Full Employment*, Basil Blackwell, Oxford and New York, 1987.

Freeman, C. and L. Soete (eds.), *New Explorations in the Economics of Technological Change*, Pinter Publishers, London and New York, 1990.

Gerybatze, A., *Innovation, Wettbewerb und Evolution*, Tubingen, 1982.

Glaister, S., Advertising Policy and Returns to Scale in Markets where Information is Passed Between Individuals, *Economica*, May 1974, pp. 139-156.

Gold, B., Technological Diffusion in Industry: Research Needs and Shortcomings, *The Journal of Industrial Economics*, vol. 29 no. 3, March 1981.

Gomulka, S., *The Theory of Technological Change and Economic Growth*, Routledge, London and New York, 1990.

Gort, M. and S. Klepper, Time Paths in the Diffusion of Product Innovations, *The Economic Journal*, vol. 92, September 1982, pp. 630-653.

Gort, M. and A. Konakayama, A Model of Diffusion in the Production of an Innovation, *The American Economic Review*, vol. 72, no. 5, December 1982,pp. 1111-1120.

Griliches, Z., Hybrid Corn: an Exploration in the Economics of Technological Change, *Econometrica*, vol 25, 1957, pp. 501-522.

Hägerstrand, T., *Innovation Diffusion as a Spatial Process*, The University of Chicago Press, Chicago, 1967.

Hannan, T.H. and J.M. McDowell, The Determinants of Technology Adoption: the Case of the Banking Firm, *Rand Journal of Economics* vol. 13, no. 3, Autumn 1984.

Hicks, J.R., *The Theory of Wages*, Macmillan, London, 1932

Intriligator, M.D., *Mathematical Optimization and Economic Theory*, Prentice Hall, Englewood Cliffs, N.J., 1971.

Ireland, N. and P. Stoneman, Technological Diffusion, Expectations and Welfare, *Oxford Economic Papers*, vol. 38, 1986.

Iwai, K., Schumpeterian Dynamics I: An Evolutionary Model of Innovation and Imitation , *Journal of Economic Behaviour and Organization*, vol. 5, 1984, pp. 159-190.

Iwai, K., Schumpeterian Dynamics II: Technological Progress, Firm Growth and 'Economic Selection', *Journal of Economic Behaviour and Organization*, vol. 5, 1984, pp. 321-351.

Jensen, R., Adoption and Diffusion of an Innovation of Uncertain Profitability, *Journal of Economic Theory*, vol. 27, 1982.

Jovanovic, B. and S. Lach, Entry, Exit, and Diffusion with Learning by Doing, *The American Economic Review*, vol. 79, no. 4, September 1989, pp. 690-699.

Kamien, M.I. and N.L. Schwartz, Optimal 'Induced' Technical Change, *Econometrica*, vol. 36, no. 1, January 1968, pp. 1-17.

Kamien, M.I. and N.L. Schwartz, Induced Factor Augmenting Technical Progress from a Microeconomic Viewpoint, *Econometrica*, vol. 37, no. 4, October 1969, pp. 668-684.

Kamien, M.I. and N.L. Schwartz, *Dynamic Optimization, The Calculus of Variations and Optimal Control in Economics and Management*, North Holland, New York, 1981.

Kamien, M.I. and N.L. Schwartz, *Market Structure and Innovation*, Cambridge University Press, Cambridge, 1982.

Karlsson, C., *Innovation Adoption and the Product Life Cycle*, Umea, 1988.

Katz, M.L. and C. Shapiro, Network Externalities, Competition, and Compatibility, *American Economic Review*, vol. 75 no. 3, 1985, pp.424-440.

Katz, M.L. and C. Shapiro, Technology Adoption in the Presence of Network Externalities, *Journal of Political Economy*, vol. 94 no. 4, 1986, pp.822-841.

Kennedy, C., Induced Bias in Innovation and the Theory of Distribution, *The Economic Journal*, vol. 74, September 1964, pp. 541-547.

Kennedy, C., A Generalization of the Theory of Induced Bias in Technical Progress, *The Economic Journal*, vol. 83, September 1973, pp. 48-57.

Kennedy C. and A.P. Thirlwall, Technical Progress: A Survey, *The Economic Journal*, vol. 82, 1972, pp. 115-176.

Kornai, J., *Anti-Equilibrium, On Economic Systems Theory and the Tasks of Research*, 1971, North-Holland.

Levin, S.G, S.L. Levin and J.B. Meisel, A Dynamic Analysis of the Adoption of a New Technology: the Case of Optical Scanners, *The Review of Economics and Statistics*, vol. 69, 1987, pp. 12-17.

Lucas, R.E., On the Mechanics of Economic Development, *Journal of Monetary Economics*, vol. 22, 1988, pp. 3-42.

Magat, W.A., Technological Advance with Depletion of Innovation Possibilities - Implications for the Dynamics of Factor Shares, *The Economic Journal*, vol. 89, September 1979, pp. 614-623.

Mahajan, V. and R.A. Peterson, *Models for Innovation Diffusion*, Sage Publications, Beverly Hills, 1985.

Mansfield, E., Technical Change and the Rate of Imitation, *Econometrica*, October 1961, pp. 741-766.

Mansfield, E., *Industrial Research and Technological Innovation*, Norton, New York, 1968.

Mathews, J., *Tools of Change: New Technology and the Democratisation of Work*, Pluto Press, Sydney, 1989.

Metcalfe, J.S., Impulse and Diffusion in the Study of Technical Change, *Futures*, vol 13, 1981.

Metcalfe, J.S., The Diffusion of Innovation : an Interpretative Survey, in : Dosi, G. *et al.* (eds.), *Technological Change and Economic Theory*, 1988.

Mohr, L., *Explaining Organizational Behaviour*, Jossey-Bass, San Fransisco, 1982.

Mueller, D.C., The Corporation and the Economist, *International Journal of Industrial Organization*, vol. 10, no. 2, 1992, pp. 147-170.

Nakicenovic, N. and A. Grübler (eds.), *Diffusion of Technologies and Social Behaviour*, IIASA and Springer Verlag, Berlin, 1991.

Nabseth, L. and G.F. Ray, *The Diffusion of New Industrial Processes: An International Study*, Cambridge University Press, Cambridge, 1974.

Nelson, R.R. and S.G. Winter, In Search of Useful Theory of Innovation, *Research Policy*, vol. 6, 1977, pp. 36-76.

Nelson, R.R. and S.G. Winter, Simulation of Schumpeterian Competition, *American Economic Review*, vol. 67, 1977, pp. 271-276.

Nelson, R.R. and S.G. Winter, The Schumpeterian Tradeoff Revisited, *American Economic Review*, vol. 72, 1982, pp. 114-132.

Nelson, R.R. and S.G. Winter, *An Evolutionary Theory of Economic Change*, Harvard University Press, Cambridge, Massachusets, 1982.

Pennings, J.M. and F. Harianto, The Diffusion of Technological Innovation in the Commercial Banking Industry, *Strategic Management Journal*, vol. 13, 1992, pp. 29-46.

Porter, M.E., *The Competitive Advantage of Nations*, Macmillan, 1990.

Porter, M.E., The Competitive Advantage of Nations, *Harvard Business Review*, March-April 1990, pp. 73-93.

Reinganum, J.F., On the Diffusion of New Technology: A Game Theoretic Approach, *Review of Economic Studies*, vol. 48, 1981.

Reinganum, J.F., Technology Adoption under Imperfect Information, *Bell Journal of Economics*, Spring 1983, pp. 57-63.

Rogers, E.M., Diffusion of Innovations, Free Press, New York, 1983.

Romer, P.M., Endogenous Technological Change, *Journal of Political Economy*, vol. 98, no. 5, pt. 2, 1990, pp. s71-s102.

Rose, N.L., and P.L. Joskow, The Diffusion of New Technologies: Evidence from the Electric Utility Industry, *Rand Journal of Economics*, vol. 21, no. 3, autumn 1990, pp. 354-373.

Rosegger, G., *The Economics of Production and Innovation*, 1986, Pergamon Press.

Rosenberg, N., Factors Affecting the Diffusion of Technology, in *Perspectives on Technology*, Cambridge Univesity Press, 1976.

Rosenberg, N., On Technological Expectations, *The Economic Journal*, vol. 86, 1976, 523-535.

Rosenberg, N., Learning by Using, in: *Inside the Black Box - Technology and Economics*, Cambridge University Press, Cambridge, 1982.

Salter, W.E.G., *Productivity and Technical Change*, Cambridge University Press, 1969.

Samuelson, P.A., A Theory of Induced Innovation along Kennedy-Weizsäcker Lines, *The Review of Economics and Statistics*, November 1965, pp. 343-356.

Salter, W.E.G., *Productivity and Technical Change*, Cambridge University Press, 1966.

Sato, R. and R. Ramachandran, Factor Price Variation and the Hicksian Hypothesis: a Microeconomic Model, *Oxford Economic Papers*, vol. 39, 1987, pp. 343-356.

Sato, R. and G. S. Suzawa, *Research and Productivity - Endogenous Technical Change*, Auburn House Publishing Company, Boston, 1983.

Schumpeter, J.A., *Business Cycles*, Porcupine Press, Philadelphia, 1939.

Schumpeter, J.A., *Capitalism, Socialism and Democracy*, George Allen and Unwin, London, 1942.

Scherer, F.M., *Industrial Market Structure and Economic Performance* (2nd ed.), Rand McNally, Chicago, 1980.

Scott, M.F., *A New View of Economic Growth*, Clarendon Press, Oxford, 1989.

Silverberg, G., Technical Progress, Capital Accumulation and Effective Demand: A Self-Organization Paradigm, in: D. Batten, J. Casti and B. Johansson (eds.), *Economic Evolution and Structural Adjustment*, Springer Verlag, Berlin, 1987.

Silverberg, G., Modelling Economic Dynamics and Technical Change: Mathematical Approaches to Self-Organisation and Evolution, in: Dosi, G. *et. al.* (eds.) *Technical Change and Economic Theory*, Pinter Publishers, London and New York, 1988.

Silverberg, G., G. Dosi and L. Orsenigo, Innovation, Diversity and Diffusion: a Self- Organisation Model, *The Economic Journal*, Dec. 1988.

Silverberg, G., Adoption and Diffusion of Technology as a Collective Evolutionary Process, in: Nakicenovic, N. and A. Grübler (eds.), *Diffusion of Technologies and Social Behaviour*, IIASA and Springer Verlag, Berlin, 1991.

Simon, H.A., On the Behavioral and Rational Foundations of Economic Dynamics, in: R.H. Day and G. Eliasson (eds.), *The Dynamics of Market Economies*, North-Holland 1986.

Soete, L. and R. Turner, Technology Diffusion and the Rate of Technical Change, *The Economic Journal*, vol. 94, 1984, pp. 612-623.

Stoneman, P., *The Economic Analysis of Technological Change*, Oxford University Press, 1983.

Stoneman, P., *The Economic Analysis of Technology Policy*, Clarendon Press, Oxford, 1987.

Stoneman, P., *Technological Diffusion The Viewpoint of Economic Theory*, paper presented at a Conference on Innovation Diffusion, Venice, March 1986.

Stoneman, P., Intra-Firm Diffusion, Bayesian Learning and Profitability, *The Economic Journal*, June 1981.

Stoneman, P. and W. Ochoro, A Means-Variance Approach to the Theory of Intrafirm Diffusion, in: T. Puu and S. Wibe (eds.), *The Economics of Technological Progress*, London, 1980.

Stoneman, P. and N.J. Ireland, The Role of Supply Factors in the Diffusion of New Process Technology, *The Economic Journal*, Supplement, vol. 93, 1983.

Thirtle, C.G. and V.W. Ruttan, *The Role of Demand and Supply in the Generation and Diffusion of Technical Change*, vol. 21 in the series : Fundamentals of Pure and Applied Economics, Harwood Academic Publishers, Chur, 1987.

De Wit, G.R., *Technologische Ontwikkelingen in het Nederlandse Bankwezen binnen een Internationale Context* (Technological Development in the Dutch Banking Industry in an International Framework), Research Memorandum 87-002, Department of Economics, University of Limburg, 1987.

De Wit, G.R., *Technologische Ontwikkelingen in het Bankwezen in vergelijking met Andere Bedrijfstakken in Nederland* (Technological Development in Banking in Comparison to Other Dutch Industries), Research Memorandum 87-007, Department of Economics, University of Limburg, 1987.

Nederlandse samenvatting - summary in Dutch

De causale relatie tussen economische en technologische ontwikkeling is tweezijdig: enerzijds bepaalt de technologische ontwikkeling de randvoorwaarden voor economische groei, anderzijds stuurt het economisch proces de ontwikkeling van nieuwe technologie. Om economische ontwikkeling te kunnen analyseren, is er daarom behoefte aan een conceptueel en wiskundig instrumentarium, toegesneden op het beschrijven en formaliseren van de rol van technologische verandering in de economie. In dit proefschrift wordt een bijdrage geleverd aan het ontwikkelen van dit instrumentarium, waarbij aansluiting wordt gezocht voor wat betreft de conceptuele en theoretische invulling bij de evolutionaire economie en voor wat betreft de wiskundige uitwerking bij de neoklassieke traditie.

Ondernemingen investeren in nieuwe produktietechnieken en veranderen daarmee hun produktiecapaciteit en hun produktiviteit, hun vraag naar arbeid, kapitaalgoederen en andere produktiemiddelen per eenheid produkt. Het centrale probleem in dit proefschrift is de wijze waarop ondernemingen keuzes maken omtrent dit soort investeringen. Het gaat hierbij om de analyse van de doelstellingen, de instrumenten en de mogelijkheden van de onderneming. Uitgangspunten zijn de veronderstelling dat de voornaamste doelstelling van de onderneming toekomstige winst is; voorts, de veronderstelling dat de mogelijkheden om winst te genereren worden beperkt door de hoedanigheid van de onderneming op het moment van keuze, door technische en marktbeperkingen en door de begrensdheid van beschikbare cognitieve capaciteiten; en tenslotte, de veronderstelling dat de werkwijze van de onderneming gekarakteriseerd kan worden als routinematig handelen, dat als gevolg van de uitkomsten van keuzeprocessen geleidelijk aan veranderende door de omgeving bepaalde voorwaarden wordt aangepast. Deze uitgangspunten worden in dit proefschrift nader uitgewerkt, geformaliseerd in modellen en getoetst aan de praktijk van het bankwezen in Nederland.

Hoofdstuk 1 geeft een duiding van het belang van verandering in technologie voor economische ontwikkeling en signaleert een toename in de belangstelling voor dit onderwerp vanuit diverse disciplines, onder andere industriële economie, groeitheorie en strategisch management. Voorts wordt gewezen op de recente ontwikkeling van een theoretische invalshoek, het 'evolutionair' perspectief, vanwaaruit bruikbare begrippen en theorieën naar voren gebracht zijn, die echter slechts sporadisch in modellen zijn uitgewerkt.

Hoofdstuk 2 levert de theoretische basis van de studie. In het eerste deel wordt gewezen op het belang van cognitieve activiteiten voor het economisch proces. Onder cognitieve activiteiten worden onder andere verstaan: het verkrijgen en beoordelen van informatie, het opdoen en ontwikkelen van kennis, het verwerven van vaardigheden, het onderbouwen van verwachtingen en inschattingen en het maken van keuzes verstaan. Er wordt betoogd dat cognitieve processen als onderdeel van economisch gedrag tot op zekere hoogte op dezelfde wijze geanalyseerd kunnen worden als de materiële produktieprocessen van een onderneming: door opofferingen van middelen worden resultaten voortgebracht. In de reële sfeer leveren arbeid en kapitaal diensten om binnen een bepaalde tijdsspanne uit grondstoffen en energie produkten te maken; in de cognitieve sfeer leveren mensen intellectuele inspanningen om in een proces van zoeken en afwegen tot informatie, kennis of beslissingen te komen. Net zoals in de materiële sfeer de produktiemogelijkheden beperkt worden door restricties van materiële aard, worden de mogelijkheden in de cognitieve sfeer beperkt door cognitieve restricties: beperkingen die voortvloeien uit het feit dat agenten (mensen danwel organisaties) tijd nodig hebben en zich inspanning moeten getroosten om informatie te verwerven en beslissingen te nemen. Wanneer het gaat om zaken van zekere complexiteit, zoals bij het ontwikkelen en invoeren van nieuwe technologieën vaak het geval is, kan de beperkte beschikbaarheid van cognitieve hulpbronnen een bindende restrictie vormen, analoog aan de schaarste van materiële hulpbronnen.

In het tweede deel van dit hoofdstuk wordt een overzicht gegeven van aspecten van het verschijnsel invoering van nieuwe technologie, waaraan in de literatuur aandacht is besteed. Aan de orde komen het verband tussen adoptie en diffusie en de vraag waar een verklaring voor de empirische regelmatigheid van de diffusiecurve gezocht moet worden: op het micro- of op het macro-economische niveau. De verklaring van een regelmatig diffusiepatroon kan terug gaan op een regelmatige verdeling van ondernemingen over een met de adoptiebeslissing samenhangende variabele, of op het feit dat adoptiebeslissingen zelf onderling op enigerlei wijze samenhangen. Verdere kwesties in dit deel betreffen het (on-)evenwichtskarakter van het diffusieproces, het endogene danwel exogene karakter van de oorzaken die het proces drijven, het routinematige aspect van economisch handelen en het discontinue karakter van technologische ontwikkeling.

Hoofdstuk 3 bouwt voort op de beschrijving van economisch gedrag in hoofdstuk 2, en brengt de voornaamste gedachten onder in een model dat het beslissingsproces van een onderneming weergeeft. De onderneming produceert op het moment van beslissing een bepaalde hoeveelheid produkt (output), met behulp van een bepaalde combinatie van middelen (inputs). Er wordt een onderscheid gemaakt tussen vaste en variabele produktiemiddelen. De doelstelling van de onderneming is maximale netto contante waarde van de toekomstige winst. De onderneming kan daartoe investeren, toevoegen aan de *vaste* produktiemiddelen, hetgeen leidt tot toename van de produktiecapaciteit (expansie) en tot verhoging van de produktiviteit van de *variabele* produktiemiddelen (rationalisatie). Enerzijds beslist de onderneming hoeveel te investeren, gegeven dat investeringen een dalend marginaal rendement hebben; anderzijds beslist de onderneming in welke verhouding te investeren in capaciteitsuitbreiding en in verhoging van de produktiviteit van de variabele middelen. Het dalend marginaal rendement op investeringen op elk moment in de tijd hangt samen met toenemende aanpassingskosten en dergelijke, maar ook met beperkingen in de cognitieve capaciteit van de onderneming.

In het model wordt geabstraheerd van technische veroudering van kapitaalgoederen, van financieringsrestricties, van markten voor tweedehands investeringsgoederen en van toe- en uittreding. Markten voor eindprodukten en produktiemiddelen ruimen voortdurend en er is geen onderbezetting. Mogelijkheden voor technologische verandering vinden hun weerslag in de technische restricties waarmee de onderneming zich geconfronteerd ziet. Indien er géén technische vooruitgang is voegen investeringen slechts meer van hetzelfde toe aan de bestaande capaciteit. Indien er wel mogelijkheden tot technische verbeteringen bestaan, komt dit tot uitdrukking in een cumulatief effect van huidige investeringen in de loop van de tijd. Dit impliceert dat niet alleen het rendement, maar ook het effect op capaciteit en produktiviteit van tegenwoordige investeringen mede bepaald wordt door toekomstige investeringen, en dat de optimaliserende onderneming derhalve een tijdpad voor de investeringen moet plannen.

De belangrijkste karakteristieken van het model kunnen als volgt worden samengevat. Ten eerste wordt aangenomen dat de onderneming noch zijn produktiecapaciteit, noch zijn produktietechniek kan wijzigen zonder investeringen. Ten tweede wordt aangenomen dat investeringen leiden tot incrementele aanpassingen van de capaciteit en incrementele verhogingen van de produktiviteit van variabele middelen. Ten derde wordt verondersteld dat zowel technische als cognitieve restricties de bewegingsvrijheid van de onderneming beperken in het aanpassen van produktieschaal en produktietechniek. Ten vierde wordt verondersteld dat, indien de positie van een onderneming op het relevante technologische traject ruimte laat voor technologische verbetering, investeringen een cumulatief effect hebben.

Dit model is allereerst gebruikt om investeringsbeslissingen van ondernemingen te analyseren. Aangetoond is dat ondernemingen die zich gedragen volgens de vooronderstellingen van het model, bij een volledig elastische vraag naar eindprodukten en een volledig elastisch aanbod van variabele produktiemiddelen, op de lange duur hun investeringen exponentieel laten groeien en hoe langer hoe meer de richting van hun investeringen verleggen van rationalisatie naar expansie. Echter, indien markten voor eindprodukten en variabele produktiemiddelen niet volledig elastisch zijn, stagneren op termijn de investeringen in expansie, ondanks het feit dat winsten positief zijn. Het investeringsgedrag van de individuele onderneming wordt cruciaal beïnvloed door de elasticiteiten van de vraag- en aanbodfuncties op de relevante markten enerzijds, en door zijn marktaandeel op die markten anderzijds.

Het model is verder gebruikt om het concurrentieproces tussen ondernemingen te onderzoeken, zowel analytisch als door middel van simulaties. De ontwikkeling van ondernemingen in een markt hangt binnen de context van dit model in hoge mate af van de startcondities: de ontwikkeling in de toekomst kan niet los gezien worden van de geschiedenis van de ondernemingen. In een pad-afhankelijke wereld ligt het niet in de rede te veronderstellen dat bij de aanvang van het proces de toestand door *stationairiteit gekarakteriseerd kan worden en is het* derhalve niet mogelijk iets algemeens over de startcondities te zeggen. De toestand van de ondernemingen bij aanvang van het concurrentieproces en de aanvankelijke verdeling van de markt kunnen vele gedaanten hebben, en afhankelijk van die begintoestand kunnen zich binnen het kader van het model legio ontwikkelingen voordoen, waarbij ondernemingen gedurende lange tijd naast elkaar voortbestaan met verschillende groeivoeten, marktaandelen, efficiëntieniveaus en snelheden van technologische ontwikkeling. Groeivoeten, marktaandelen en winsten kunnen, afhankelijk van het ontwikkelingsstadium van de afzetmarkt, divergeren of convergeren. Een voorbeeld hiervan, een markt met zes ondernemingen van verschillende omvang en efficiëntie, is met de hulp van computersimulaties uitgewerkt.

De computersimulaties betreffen doorrekeningen van marktinteracties tussen ondernemingen die zich gedragen volgens het microeconomische investeringsmodel. Een van de uitkomsten van de computersimulaties is dat een aantal geaggregeerde variabelen zich blijken te gedragen volgens het patronen die bekend zijn uit de literatuur over diffusie en over de produkt-levenscyclus. In de loop van de tijd blijkt de afzet toe te nemen volgens een sigmoïde patroon, terwijl ten gevolge van procesinnovaties de vraag naar variabele produktiemiddelen weinig variatie vertoont. Prijsontwikkelingen weerspiegelen volume-ontwikkelingen: de afzetprijs daalt scherp, terwijl de prijs van variabele produktiemiddelen weinig variëert. De investeringen stijgen aanvankelijk en zijn in de eerste fase van de ontwikkeling van de markt gericht op uitbreiding van de produktiecapaciteit. Later dalen de investeringen en verschuift het accent naar verhoging van de produktiviteit. De grootste winsten worden gemaakt in de meest expansieve fase van de marktontwikkeling, voordat de markt zijn grootste omvang bereikt heeft.

Het verband tussen marktstructuur en de snelheid van technologische ontwikkeling blijkt nogal complex in dit model vanwege de simultane afhankelijkheid van investeringsgedrag van prijselasticiteiten, marktaandelen en ondernemingsomvang. Een grote onderneming profiteert van schaalvoordelen bij het genereren en implementeren van nieuwe technologie, hetgeen samenhangt met het cumulatieve karakter van technologische ontwikkeling, en zal dus geneigd zijn innovatiever te zijn dan een kleine onderneming. Daar staat tegenover, dat een grote onderneming vaak een groot marktaandeel heeft, en daardoor de daling van de afzetprijs, die het gevolg is van een procentuele vergroting van de afzet, in meerdere mate internaliseert dan een onderneming met een klein marktaandeel. Een groot marktaandeel zal daarom investeringen in nieuwe technologie, indien die leiden tot expansie, afremmen. In de loop van de tijd, als de afzetmarkt verzadigd raakt, verandert het samenspel van de factoren ondernemingsomvang en marktaandeel bij de bepaling van de omvang van de investeringen in capaciteitsuitbreiding. Dit verklaart het optreden van een sigmoïde diffusiecurve van het eindprodukt binnen dit model. In tegenstelling tot expansieve investeringen, worden investeringen in verhoging van de produktiviteit gestimuleerd door een groot marktaandeel. Hoezeer een groot marktaandeel de expansie van een onderneming afremt en de rationalisatie stimuleert, hangt af van de prijselasticiteiten van de vraag naar eindprodukten en het aanbod van variabele produktiemiddelen.

Hoofdstuk 4 vormt de brug tussen de theoretische uiteenzettingen en de empirische toetsing. Om empirische ondersteuning te vinden voor de eerder geschetste modelmatige benaderingen zijn data gebruikt die stammen van het Nederlandse bankwezen. Het bankwezen vormt een substantiële sector in de Nederlandse economie, waarin een aantal recente trends die algemeen voorkomen in dienstensectoren in westerse economieën, prominent naar voren treden. Enerzijds groeien dienstensectoren in belang, relatief ten opzichte van industriële sectoren, anderzijds ontwikkelen ze meer en meer karakteristieken die voorheen kenmerkend waren voor industriële sectoren: diensten worden kapitaalintensiever, produkten worden meer en meer gestandaardiseerd en technologie speelt een steeds voornamere rol. Een belangrijke oorzaak van deze trend is de ontwikkeling van computer- en telecommunicatietechnologie. Het bankwezen is een duidelijk voorbeeld van deze ontwikkeling: het heeft op grote schaal computer en telecommunicatie apparatuur in gebruik genomen, daarmee de produktiekosten drastisch teruggebracht en de dienstverlening en het aanbod van produkten behoorlijk uitgebreid.

Voor de toetsing van de modellen was een dataverzameling beschikbaar met cijfers van kantoren van een grote Nederlandse coöperative bank. In dit hoofdstuk worden deze cijfers inductief geanalyseerd. Het blijkt onder meer dat de diffusie van nieuwe technieken in het kantorennet in een aantal gevallen een sigmoïde curve doorloopt, en dat er sprake is van schaalvoordelen in produktie. Voor het testen van de modellen uit het vorige hoofdstuk worden variabelen geselecteerd die respectievelijk produktiecapaciteit, variabele produktiemiddelen, investeringen en technologisch niveau benaderen. De voornaamste investeringen in technologische ontwikkeling in banken op lokaal niveau in de periode 1979 tot 1987 werden gedaan in automatische gegevensverwerking. Allereerst werd er geïnvesteerd in automatisering van de 'back office', daarna in automatisering van de 'front office', en vervolgens in het installeren van geldautomaten en daarmee samenhangende communicatienetwerken. Het belangrijkste gestandaardiseerde massaprodukt van deze lokale banken, waarop het grootste deel van de automatiseringsinspanningen gericht was, is de rekeningenadministratie, met inbegrip van de administratie van overschrijvingen, bij- en afboekingen. Hoewel het proces tegenwoordig relatief kapitaalintensief is, behoren verschillende soorten arbeid, te weten baliediensten, dataverwerking, administratie en management, toch nog tot de voornaamste produktiefactoren in het produktieproces. Geen van deze soorten arbeid zijn volledig als variabele produktiemiddelen te karakteriseren, maar wellicht is arbeid op de korte termijn meer variabel dan andere produktiemiddelen. Derhalve zijn data betreffende mutaties in de rekeningenadministratie gekozen als benadering voor produktiecapaciteit, data betreffende verschillende typen arbeid als benadering voor variabele produktiemiddelen, over een aantal jaren gecumuleerde investeringen in automatisering als benadering voor investeringen en de beschikbaarheid van drie verschillende automatiseringssystemen als technologie-indicator.

Hoofdstuk 5 is gewijd aan de toetsing van de modellen uit hoofdstuk 3. Omdat de beschikbare data voornamelijk dwarsdoorsnede data zijn, terwijl volledige toetsing van het model tijdreeksen zou vereisen, is de toetsing beperkt tot een aantal aspecten. Het model is vereenvoudigd tot een twee-periodenmodel en er is met name gekeken naar de robustheid van parameters bij beschouwing van verschillende banken: er is gekeken in welke mate het investeringsgedrag van banken weergegeven kan worden met behulp van dezelfde modelparameters. Een eerste vraag die aan de orde komt is in hoeverre investeringen een cumulatief effect hebben op produktie en op de vraag naar variabele produktiemiddelen, in hoeverre er aanwijzingen zijn dat er een proces van technologische vooruitgang plaatsvindt. Een tweede vraag is in hoeverre technologische mogelijkheden en technologische ontwikkeling van verschillende banken op verschillende momenten door eenzelfde investeringsmodel kunnen worden weergegeven, in hoeverre ondernemingen geconfronteerd worden met gelijke technische beperkingen.

De resultaten van de modelschattingen dienen met omzichtigheid geïnterpreteerd te worden, omdat het model sterk gestyleerd is en de aansluiting tussen de data en de modelvariabelen te wensen overlaat. Niettemin dringen een aantal conclusies zich op. Schattingsresultaten wijzen erop dat banken scherp dalende marginale rendementen op investeringen ondervinden op elk specifiek moment in de tijd. Desalniettemin investeren grote banken meer dan kleine. Parameterschattingen suggereren dat dit verklaard kan worden door het feit dat grote banken bij investeringen kunnen profiteren van schaalvoordelen. In absolute termen levert dezelfde investering een grote bank meer extra capaciteit op dan een kleine bank, en daarom kan een grote bank een groter budget investeren voordat marginale opbrengsten aan marginale kosten gelijk zijn. Deze uitkomsten zijn verenigbaar met de hypothese dat investeringen een cumulatief effect hebben, hetgeen erop kan wijzen dat investeringen de produktiviteit van het produktie-apparaat omhoog brengen. Het verband tussen investeringen en de vraag naar variabele produktiemiddelen bleek minder eenduidig te zijn, maar ook hier kan een cumulatief effect niet worden uitgesloten.

Betreffende de vragen in hoeverre banken geconfronteerd worden met dezelfde technologische restricties en in welke mate ze hetzelfde technologische traject volgen, konden slechts voorlopige antwoorden gevonden worden. Hoewel alle banken in kwalitatieve zin dezelfde technologische stappen zetten, van 'back office' automatisering, via 'front office' automatisering, naar installatie van geldautomaten, lijken ze dit pad niet met dezelfde snelheid of in dezelfde volgorde af te lopen. Banken investeren niet in dezelfde richting (dezelfde verhouding expansie-rationalisatie); ze investeren voorts niet zozeer vergelijkbare absolute bedragen in automatisering alswel eenzelfde deel van de omzet.

Hoofdstuk 6 begint met de constatering dat uit hoofdstuk 4 naar voren gekomen is dat diffusiepatronen van produktietechnologieën in banken een zekere mate van regelmaat te zien geven, maar dat in hoofdstuk 5 niet afdoende is aangetoond dat dit verklaard kan worden op grond van kosten en baten van het invoeren van technieken, voor zover deze uit de cijfers blijken. Dit werpt de vraag op die in hoofdstuk 2 al aan de orde is geweest: waar bevindt zich het mechanisme dat leidt tot de regulariteit van het diffusiepatroon, op het microniveau of op het niveau van de markt of de sector? Indien kosten en baten van het invoeren van nieuwe technologieën moeilijk kwantificeerbaar zijn en informatie over nieuwe technieken schaars is, kunnen 'bandwagon' effecten een voorname rol gaan spelen bij de verspreiding van een innovatie: ondernemingen vertonen imitatief gedrag, in een streven naar risicomijding. Een verwante hypothese is dat bij het ontbreken van eenduidige informatie over kosten en baten van een innovatie gedrag van ondernemingen in zekere mate erratisch wordt, dat in de markt een selectieproces bepaalt welke ondernemingen groeien en welke ten onder gaan, en dat dit selectieproces een ordelijk patroon op geaggregeerd niveau tot gevolg heeft. Deze noties zijn in hoofdstuk 6 in modellen van overgangskansen uitgewerkt. Hierbij wordt ervan uitgegaan dat de relatieve (verwachte) winstgevendheid van een techniek en de relatieve frequentie van het gebruik de kans bepalen dat een onderneming van een minder winstgevende techniek naar een meer winstgevende overspringt. Er wordt een onderscheid gemaakt tussen twee mechanismen, een 'duw'- en een 'trek'- mechanisme. Het eerste mechanisme werkt als een onderneming zich onder druk van de concurrentie gedwongen voelt om een oude techniek op te geven, en vervolgens naar een nieuwe techniek op zoek gaat; het tweede mechanisme werkt als een onderneming aangetrokken wordt door een nieuwe techniek en daarom de oude opgeeft. De twee verschillende specificaties leiden tot twee verschillende diffusiecurves, waarbij het duwmodel een sigmoïde curve oplevert en het trekmodel een concave curve. Het ligt in de lijn der verwachting dat het duwmechanisme werkt bij relatief dure en complexe innovaties, terwijl het trekmechanisme meer van belang is bij relatief goedkope en ongecompliceerde innovaties.

Beide modellen zijn geschat met de data betreffende invoering van nieuwe technologieën in banken uit hoofdstuk 4. Het bleek dat in dit geval een model met een duwmechanisme beter voldoet dan een model met een trekmechanisme, een uitkomst geheel in de lijn der verwachting, gegeven het complexe karakter van de innovaties hier aan de orde. Een vergelijking van de uitkomsten van empirische toetsen in hoofdstukken 5 en 6 suggereert dat, benevens het streven naar capaciteits- en produktiviteitsgroei, mechanismen op geaggregeerd niveau zoals imitatie, 'bandwagon' effecten en spill-over mechanismen een rol hebben gespeeld bij de diffusie van produktietechnieken in banken.

***Hoofdstuk** 7* besluit met een samenvatting, een opsomming van de voornaamste conclusies en een korte vooruitblik.

Curriculum Vitae

Paul Diederen is geboren op 9 september 1959 te Brunssum. In 1977 heeft hij met goed gevolg het Gymnasium-β afgesloten aan het St. Janscollege te Hoensbroek. Van 1977 tot 1978 studeerde hij aan Muhlenberg College, Allentown, Pennsylvania. Van 1978 tot 1986 studeerde hij econometrie aan de Universiteit van Amsterdam en van 1981 tot 1983 tevens wijsbegeerte. Tijdens zijn studie was hij werkzaam als student-assistent bij de vakgroep macroeconomie aan de Faculteit der Economische Wetenschappen van de Universiteit van Amsterdam. Na het behalen van het doctoraaldiploma econometrie trad hij medio 1986 in dienst bij de Faculteit der Economische Wetenschappen van de Rijksuniversiteit Limburg te Maastricht. Aanvankelijk als wetenschappelijk assistent bij de vakgroep kwantitatieve economie, later als toegevoegd docent en toegevoegd onderzoeker, verrichtte hij onderzoek naar diffusie van innovaties. Vanaf 1987 is hij verbonden geweest aan het onderzoeksinstituut MERIT. Momenteel werkt hij aan het Warwick Business School Research Bureau, University of Warwick, Coventry, Verenigd Koninkrijk.

Paul Diederen (born September 9, 1959 in Brunssum, The Netherlands) studied econometrics at the University of Amsterdam from 1978 to 1986. He worked as a research assistent with the chair of macroeconomics at the Department of Economics of Amsterdam University. Between 1986 and 1993 he held posts as research fellow and as assistant professor at MERIT and at the Department of Economics at the University of Limburg in Maastricht. Presently he has a position at the Warwick Business School Research Bureau, University of Warwick, Coventry, United Kingdom.

Maastricht Economic Research Institute
on Innovation and Technology

Datawyse | Universitaire Pers Maastricht
ISBN 90 5278 067 6
NUGI 681