Kuussaari, Harri

Working Paper

Productive efficiency in Finnish local banking during 1985-1990

Bank of Finland Discussion Papers, No. 14/1993

Provided in Cooperation with:
Bank of Finland, Helsinki

Suggested Citation: Kuussaari, Harri (1993) : Productive efficiency in Finnish local banking during 1985-1990, Bank of Finland Discussion Papers, No. 14/1993, ISBN 951-686-380-9, Bank of Finland, Helsinki,
https://nbn-resolving.de/urn:nbn:fi:bof-201908121394

This Version is available at:
https://hdl.handle.net/10419/211686

BANK OF FINLAND DISCUSSION PAPERS **14/93**

Harri Kuussaari*

Research Department
5.11.1993

Productive Efficiency in Finnish Local Banking During 1985–1990

* I am greatly indebted to Jouko Vilmunen for much valuable advice. I would also like to thank Juha Tarkka and Heikki Koskenkylä for helpful comments and suggestions.

ISBN 951-686-380-9
ISSN 0785-3572

Suomen Pankin monistuskeskus
Helsinki 1993

Abstract

The study is concerned with measuring and explaining producer performance in Finnish local banking. Primary interest is in the production of retail banking services. The main objective is to find out how large and what kind of relative efficiency differences has there been in Finnish local banking. The overall productive efficiency will be decomposed into technical efficiency and scale efficiency. Technical inefficiencies are generally a result of wasteful use of inputs which is due to differences in managerial abilities to control costs and maximize revenues. Scale inefficiency on its half occurs when a bank's size is not optimal i.e. average costs are not minimized. Total productivity of local banks is studied with Malmquist-indices. Productivity growth will be decomposed into effects of technological advancement and to change of technical efficiency. One purpose of this study is to empirically look at the implications of bank mergers for efficiency and productivity. Time period under study is from 1985 to 1990. The data includes all Finnish cooperative and savings banks. Relatively efficient production frontier is constructed using a non-parametric linear programming method called data envelopment analysis (DEA).

The main result of the efficiency analysis is that technical inefficiencies dominate scale inefficiencies in Finnish local banking. Distributions of technical inefficiencies are similar for cooperative and savings banks, but savings banks were found to be slightly more scale inefficient. Since it is possible that an average bank is continuously about 20 % more inefficient than the best practise banks, it appears that there is not much competitive pressure to control costs. The large and persistent cost efficiency differences between banks of similar size and product mix suggest that greater competition within the banking industry would be beneficial. The main result of the productivity analysis is that average productivity growth was found to be totally due to technological progress rather than improvements in relative efficiency.

Tiivistelmä

Tutkimuksen tarkoituksena on selvittää suomalaisten paikallispankkien palvelutuotannon tehokkuutta ja tuottavuuden muutoksia vuosina 1985–1990. Käytetty aineisto sisältää kaikki kyseisinä vuosina toimineet osuus- ja säästöpankit. Tarkastelu keskittyy vähittäispankkipalveluiden, erityisesti lainojen ja talletusten tuotantoon, mutta myös muun tyyppisiä pankkipalveluita on pyritty ottamaan huomioon. Tutkimuksen keskeisin tavoite on selvittää kuinka suuria suhteellisia kustannustehokkuuseroja paikallispankkien välillä on. Pankkien tuotannon tehokkuus on jaettu kahteen tekijään: pankin koosta johtuvaan skaalatehokkuuteen ja pankin sisäiseen resurssien käytön tehokkuuteen, jota kutsutaan tekniseksi tehokkuudeksi. Menetelmänä tehokkuustarkastelussa on käytetty lineaariseen ohjelmointiin perustuvaa Data Envelopment Analysis (DEA) menetelmää. Tuottavuutta on tarkasteltu Malmquist-indeksin avulla. Tuottavuuden muutokset on jaettu suhteellisen tehokkuuden muutoksen ja teknologisen kehityksen komponentteihin.

Tehokkuustarkastelun keskeisin tulos on, että pankkien kustannustehokkuus riippuu ensisijaisesti teknisestä tehokkuudesta. Pankin koon merkitys tehokkuuden kannalta on huomattavasti pienempi. Keskimääräinen paikallispankki oli noin 20 % tehottomampi kuin suhteellisesti tehokkaat paikallispankit. Keskimääräinen tehottomuus pysyi samalla tasolla ajanjakson jokaisen vuoden poikkileikkaus tarkastelussa. Tämä indikoi, että kilpailulliset paineet kustannusten tiukkaan kontrollointiin eivät olleet kovin suuria vuosien 1985–1990 aikana. Tuottavuus tarkastelun keskeisin tulos oli, että paikallispankkisektorin keskimääräinen tuottavuuden kasvu johtui täysin teknologisesta kehityksestä suhteellisen tehokkuuden pysyessä keskimäärin muuttumattomana. Osuus- ja säästöpankkien välillä ei ollut merkittäviä eroja teknisessä tehokkuudessa, mutta säästöpankit olivat keskimäärin hieman skaalatehottomampia kuin osuuspankit. On kuitenkin huomattava, että tämän hetken säästöpankkien suuret luottotappiot aiheuttanut luottoekspansio tapahtui juuri 1980-luvun lopulla. Koska tappiot eivät kuitenkaan ehtineet realisoitua tarkasteltavan ajanjakson aikana, niin erityisesti säästöpankkien tuottavuuden kasvuun ja tehokkuustuloksiin on suhtauduttava tämä varaus mielessä.

Contents

1 Introduction

The changing nature of competitive environment and regulation in banking industry has escalated empirical and theoretical research on the performance of the banks. Studies on efficiency and productivity have been especially numerous in the recent years. Bank's ability to operate efficiently is currently more crucial for bank's profitability than it was during regulation. If profound productive inefficiencies exist in banking industry, it may have to go through significant structural changes in order to adjust to the new environment. By studying the performance of the banks one can gain some insight of what kind of banks are able to survive over time and how banks are reacting to new environment. Also, studying the effects of changing environment on the banks is very important for the regulative policy purposes.

Productive efficiency of the Finnish banking sector has not been studied extensively before. This study attempts to correct some of this shortcoming. The main objective is to find out how large and what kind of relative efficiency differences has there been in Finnish local banking. The overall productive efficiency will be decomposed into technical and scale efficiency. Technical inefficiencies are generally a result of wasteful use of inputs which is due to differences in managerial abilities to control costs and maximize revenues. Scale inefficiency on its half occurs when a bank's size is not optimal i.e. average costs are not minimized. Whether the main source of inefficiency is technical in nature or depends on the scale of operations, has different implications e.g. for competitive environment, market structure, and policy purposes. Development of productive efficiency is studied in a period from 1985 to 1990. The data includes all Finnish cooperative and savings banks. If banks are generally becoming more efficient, then better profitability and lower prices could be expected. Total productivity of local banks is studied with Malmquist-indices. Development of productivity growth in banking during the years 1985–1990 will be decomposed into effects of technological advancement and to change of technical efficiency.

A number of local banks have merged with one another claiming that increased technical and scale efficiency will result in reduced costs. If mergers truly increase technical and scale efficiency, then it could be expected that possible inefficiencies would be wrung out of the industry. Market forces would drive the inefficient banks either to merge or exit the industry. On the other hand, the primary motives of the mergers in Finnish local banking at the later half of 1980's might have been different from increasing efficiency. One goal of this study is to empirically look at the implications of these bank mergers for efficiency and productivity.

There is no consensus on the best method for estimating inefficiencies of the banking industry. Because efficient production technology in banking industry is unknown, the reference technology must be constructed from the available data. Observations are then compared relative to the generated production or cost frontier. Thus, all the efficiency measures are always only relative measures determined by the best practise frontier, which does not necessarily characterize an absolutely efficient production technology. In this study a non-parametric linear programming approach is chosen for the construction of the relatively efficient production frontier.

This study is organized as follows. Chapter 2 shortly describes the evolution of banking efficiency research and reviews the alternative measurement approaches that have appeared in the recent banking efficiency literature. Chapter 3 defines the concepts of technical, allocative and scale efficiency and measures for them. A linear programming method for calculating efficiency estimates is presented in chapter 4. Chapter 5 discusses the alternative ways to model bank production. The results of empirical study are reviewed in the chapter 6. Conclusions and final remarks are drawn in chapter 7.

2 Review of the literature

Until the end of 1980's, studies on efficiency of banking industry focused mainly on scale and scope economies. Most of the studies assumed at least implicitly that productive efficiency was equal throughout the banking sector, which means that all banks have equally good management and equally skilled personnel. Thus, if two banks have the same amount of resources at their use, then the banks will produce exactly the same amount of outputs. Common result of these studies was that scale and scope economies account only for few percentages of bank costs.[1] These results must, however, be viewed critically because of their ignorance of possible productive efficiency differences among banks. For example, bank's rapid growth in size might be due to its competitive advantage resulting from good productive efficiency rather than existence of scale economies. These kind of results were supported by Berger and Humphrey (1991), who found that operational inefficiencies dominate scale and product mix economies in banking. Their research also concluded that most of the inefficiency raises from overuse of physical inputs (technical inefficiency) rather than from improper mix of inputs (allocative inefficiency). Another shortcoming of the scale economies literature is that scale economies is a local concept that may not adequately capture the global aspects of scale efficiencies. Scale economies measure the cost savings from changing output marginally, but they do not account for the full benefits of moving all the way to the minimum average cost point (Berger 1992).

Recently, equal efficiency assumption has been relaxed and attention has been focused on estimating the differences in efficiencies across the banks. The choice of efficiency estimation method depends on the way that production technology is described. If duality conditions are satisfied production technology can be characterized either with production, cost or profit function (see section 3.1). Most frequent in the literature have been estimations of best practise i.e. frontier production and cost functions. Mostly interest has been focused on the shape of the efficient frontier and especially on the magnitude of deviations below the frontier. Two major approaches for constructing the relatively efficient frontiers are the econometric approaches (cost frontiers) and the linear programming approach called data envelopment analysis (production frontiers). The third approach uses profit function, but its empirical applications are still very few. These three alternative efficiency evaluation techniques are quite different from each other and all have advantages as well as disadvantages. Majority of the empirical applications are from the past few years and most of them have concentrated on U.S. banking industry.

Three different types of econometric approaches have been used for estimating technical inefficiencies. The first one is econometric frontier approach. It uses modified cost function to reveal possible inefficiencies. The general idea of using cost function is that producer's goal is to produce any amount of outputs with minimum costs. However, it is possible that he fails in this task because of productive inefficiency. Cost function characterizes the

[1] For review of the literature of scale and scope economies in banking see e.g. Kolari, Zardkoohi (1987), Humphrey (1990) and Forestieri (1993).

frontier and inefficiencies are included in the error term, which can be decomposed into technical inefficiency, allocative inefficiency and random noise. Inefficiencies are disentangled from random fluctuations by using different distributional assumptions. It is usually assumed that inefficiencies follow an asymmetric half-normal distribution, while random fluctuations follow a symmetric normal distribution.

Econometric frontier approach has been applied to banking by Ferrier and Lovell (1990) and by Bauer et al. (1993). Ferrier and Lovell found that average bank's operating costs were about 26 % higher than frontier costs. This inefficiency was decomposed into 9 % of technical inefficiency and 17 % of allocative inefficiency. The high figure of allocative inefficiency was mainly due to overuse of labour relative to use of capital and other materials. The study also concluded that there are potential cost advantages due to scale economies, but large banks did not, however, appear more cost efficient than small ones.

Berger and Humphrey (1991) used a different econometric approach, so called thick frontier approach. The idea of thick frontier is that instead of trying to estimate a precise frontier edge, cost functions are estimated for the lowest and highest average cost quartiles of banks. Banks in the lowest cost quartile are assumed to represent banks with greater than average efficiency. Cost function estimated using data from banks in the highest cost quartile represents banks with less than average efficiency. Error terms within the lowest and highest cost quartiles reflect only random error, while the differences between the quartiles reflect inefficiencies. A benefit of thick frontier approach is that it requires less specific statistical assumptions than traditional econometric approach. Berger and Humphrey (1991) found total operating inefficiency to be about 25 % of costs. In contrast to Ferrier and Lovell's (1990) econometric results Berger and Humphrey found technical inefficiencies to be greater than allocative inefficiencies. Thick frontier approach has been applied to banking also by Berger and Humphrey (1990) and by Bauer et al. (1993).

The third type of econometric approach to efficiency measurement has been presented by Berger (1992). He used distribution free approach, which avoids arbitrary distributional assumptions on the inefficiencies. The typical half-normal assumption on the inefficiencies can be ignored by assuming that random error averages out over time and inefficiencies are persistent. Berger used U.S. banking data for years 1980–1989 and found similar results as other studies, i.e. technical inefficiencies or managerial differences in banking are important and dominate the effects of scale efficiency differences. Another important result of the study was that distributional assumptions usually imposed in the literature were not consistent with the data used in the study. Distribution free approach has been used also in the banking studies by Berger (1991) and Berger and Humphrey (1992).

In the recent efficiency literature a linear programming approach called data envelopment analysis (DEA) has become a popular alternative to econometric methods. DEA is a non-parametric method for constructing deterministic frontiers. DEA is able to evaluate the relative efficiency of a set of organisations, which use multiple inputs to produce multiple outputs, while the efficient production technology is unknown. The relatively efficient frontier is determined by comparing the observed input-output combinations of the units in the sample. Number of applications using DEA to different real-world

problems has increased rapidly. Variation of the applications is also very wide. It has even been argued that DEA merits consideration as a primary method for measuring and partitioning overall technical inefficiency (see Liebenstein and Maital, 1992)

In the past few years DEA has also been frequently applied to banking industry. The first applications analyzed efficiencies of different branches of a single bank. Sherman and Gold (1985) analyzed the overall efficiency of 14 branches of a U.S. savings bank. DEA results showed that six branches were operating inefficiently compared to the others. Parkan's (1987) similar study suggested that eleven branches out of thirty-five were relatively inefficient. The samples in these studies were however quite small so that some of DEA's large sample discriminatory power could have been lost.

Rangan et al. (1988) used DEA to analyze 215 independent U.S. banks. The purpose of the study was to measure technical and scale inefficiencies of the sampled banks. The results indicated that banks could have produced the same level of output with only 70 % of the inputs actually used, while scale inefficiencies of the banks were relatively small. Aly et al. (1990) extended the work of Rangan et al. (1988) to include allocative inefficiencies and to determine whether there are differences in efficiency between unit and branch banking organisation forms. Although these two organisational forms are very different in legal environments, no significant differences in overall efficiency were found. Main source of inefficiency was technical in nature, rather than allocative or due to scale effects.

Elyasiani and Mehdian (1990) studied bank efficiency and the rate of technological change for large U.S. banks. Unlike most of the DEA studies they used intermediation approach to output measurement.[2] The results showed that the efficient frontier had shifted inward between years 1980 and 1985 due to technological advancement. Banks that had been fully efficient in 1980, on the average, could have produced the same amount of output with 90 % of the inputs, if the 1985 technology had been available. Elyaisiani and Mehdian (1992) studied possible efficiency differences of minority and non-minority owned banks with similar input-output setting. No clear ownership-efficiency relationship was found and the study reported an average overall inefficiency to be a little over 10 %.

The efficiency of Norwegian banking industry has been studied by Berg et al. (1989, 1992, 1993) and Berg (1992). Berg et al. (1989) demonstrate that the way the bank output is chosen is critically important for the identification of inefficient banks. Two alternative output definitions in the production approach are to measure loans and deposits by money values or by number of accounts. Both output identifications yielded similar distributions for efficiency scores, but rankings between banks were distinctly different, depending on the way output activity was measured. Berg's (1992) research using DEA concluded that mergers did not appear to have any significant effect on the efficiency of merging banks. Also Berger and Humphrey (1992) arrived to the conclusion that mergers did not result in significant cost efficiency gains on average,

[2] Intermediation approach treats deposits as inputs, while in the production approach deposits are considered to be outputs. Bank output measurement will be discussed in more detail in section 5.

although some mergers were very successful. Berger and Humphrey used distribution free approach for efficiency measurement.

Some comparisons between the econometric and linear programming approaches have been made. In the study by Ferrier and Lovell (1990), empirical results showed that both techniques yield similar outcomes for cost economies but dissimilar for cost efficiencies. Linear programming method resulted on average 17 % technical and 5 % allocative inefficiencies. For the econometric approach percentages were 9 and 17 respectively. Ferrier and Lovell suggest four factors that might lead to the lack of harmony between the two sets of efficiency results. First, since programming approach is not stochastic, it interprets noise as inefficiency. Second, the econometric approach imposes parametric structure on both technology and the distribution of inefficiency, and so commingles specification error with inefficiency. Third, econometric approach can determine allocative efficiency only as a mean value over the sample, so the true values of allocative efficiencies are always either overstated or understated. Fourth, in the linear programming case the variables indicating institutional type of banks were excluded to avoid excessive categorisation. Two of these categorical variables however had statistically significant impact on costs in the econometric approach.

Almost all efficiency studies have been concerned with inefficient use of inputs. A recent exception is a study by Berger, Hancock and Humphrey (1993), which uses a profit function instead of cost function to obtain efficiency measures for U.S. banks. With profit function it is possible, in addition to input inefficiencies, to study also inefficiencies of the output side of the bank. Profit function takes into account the revenue effects of producing wrong level or mix of outputs. The idea is that bank's input-output decisions may be based on shadow relative prices that differ from actual relative prices. Bank's production plan can therefore be suboptimal and lead to allocative inefficiency. Technical inefficiency is measured as deviation from desired input-output level determined by the shadow relative prices. These measures of inefficiency are thus different from radial measures used in econometric and DEA-approaches. The result of the study by Berger, Hancock and Humphrey (1993) was that most inefficiencies were due to deficient output revenues, rather than excessive input costs. They also found that large banks were more efficient than small banks, which may offset scale diseconomies found in other studies.

The major issues in measuring inefficiencies in banking concern the identification of outputs and assumptions about distributions of inefficiencies and random errors. Although there is no consensus on exactly how large technical inefficiencies in banking are, there is a consensus that they are substantial and that they dominate allocative and scale inefficiencies. The most common outcome of the efficiency studies is that average technical inefficiency is about 20 % of bank's costs. The results of DEA studies vary a little more than the results of econometric approaches. While the average levels of inefficiencies are somewhat in line between different approaches, the same is not true for the rankings of individual banks. The result of efficiency evaluation of individual bank seems to depend on the applied methodology, and therefore very strong bank specific conclusions are not necessarily reliable.

3 Efficiency measurement

Basic concept of efficiency is very simple. Efficiency measures how well a producer succeeds in transforming inputs into outputs according to his behavioural objectives. Producer is said to be efficient if he is able to achieve his goals and inefficient if he fails. Usually producer's goal is assumed to be cost minimization of production i.e. producer tries to reach economic efficiency. Any waste of inputs is to be avoided so that there is no idleness or functionless use of resources.

In production theory it is often assumed that producers are behaving efficiently in an economic sense. That means they are able to successfully allocate all resources in an efficient manner relative to the constraints imposed by the structure of production technology and by the structure of input and output markets, and relative to whatever behavioural goals are attributed to the producers (Färe et al. 1985, 5). Although economic efficiency is an obvious goal for a firm that wishes to maximize its profits, in reality it is rare that a firm could totally utilize all its resources at the maximum level.

Early studies of efficiency and its measurement were quite heterogeneous. Wide variety of models were set up to investigate a wide range of efficiency related issues in a wide range of environments. Koopmans (1951, 60) provided a formal definition of technical efficiency: a producer is technically efficient if an increase in any output requires a reduction in at least one other output or an increase in at least one input, and if a reduction in any input requires an increase in at least one other input or a reduction in at least one output. Debreu (1951) was the first to provide a measure or index of the degree of technical efficiency. His 'coefficient of resource utilization' represented the smallest fraction of the actually needed resources that would be enough for the production of certain output level. Farrell's (1957) article, however, has been the most influential paper in the efficiency measurement literature. Farrell decomposed overall efficiency into technical and allocative components and proposed indices for these measures. Farrell's efficiency measures were valid for restrictive technologies, but did not generalize easily to technologies that are not linearly homogeneous or to technologies in which strong input disposability and strict quasiconcavity are impropriate (Førsund et al. 1980). Generalizations of Farrell's efficiency measures have later been presented by Färe and Lovell (1978) and Førsund and Hjalmarsson (1974, 1979).

3.1 Alternative descriptions of the production technology

Assume that a firm uses n inputs $x \equiv (x_1,...,x_n)'$ for the production of single output y. The prices of inputs are fixed $w \equiv (w_1,...,w_n)' > 0$ and the output can be sold at a fixed price $p > 0$. Production function $f(x)$ characterizes the efficient transformation of inputs into outputs i.e. tells the maximum output achievable from the inputs available. An equivalent way of describing production technology is to use cost function, assuming that duality conditions are satisfied. The cost function $c(y,w) \equiv \min_x \{w'x \mid f(x) \geq y, x \geq 0\}$ shows the minimum costs required for different output levels. Cost minimizing input

demands can be obtained from the Shephard's lemma, $x_i(w,y) = \partial c(w,y)/\partial w_i$. Under certain regularity conditions an efficient production technology can equivalently be presented by a profit function $\pi(p,w) \equiv \max_y \{py - w'x \mid f(x) \geq y,\ x \geq 0,\ y \geq 0\}$, which shows the maximum profit available given output price p and input prices w. Profit maximizing output supply and input demands can be obtained from Hotelling's lemma, $y(p,w) = \partial\pi(p,w)/\partial p$ and $x_i(p,w) = -\partial\pi(p,w)/\partial w_i$.

3.2 Technical efficiency

Suppose that the firm's observed input-output combination is $(x^0,\ y^0)$. Production of the firm is said to be technically efficient if $y^0 = f(x^0)$ and technically inefficient if $y^0 < f(x^0)$ ($y^0 > f(x^0)$ is assumed to be impossible). Technical efficiency can be measured as a ratio between the actual output produced by the firm and the maximal output defined by the production function. That is

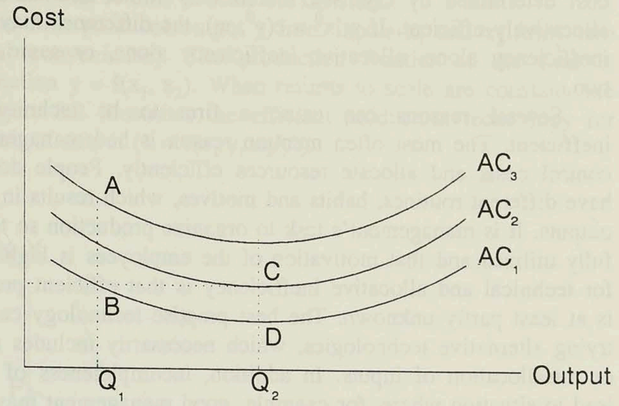
$$0 \leq \frac{y^0}{f(x^0)} \leq 1 \qquad (1)$$

Technical inefficiency results from wasteful overuse of inputs and causes therefore extra costs.

Figure 3.1 illustrates the effect that technical inefficiency has on average costs. The lowest average cost curve (AC_1) shows costs of efficient production technology. If the firm uses more inputs than necessary, then the firm's average cost curve is on higher level (AC_2 or AC_3).[3] Cost efficiency differences between firms are two-fold. For example at output level Q_1 firm B is technically efficient but firm A is inefficient. Firm C produces more output than firm B with same average costs. C is however not technically efficient, since its average costs are higher than firm D's, which produces the same level of output as C. Cost advantage of the firm D compared to firm B is due to scale economies. Thus, efficiency differences at fixed output level are called technical inefficiencies and average cost differences on the same AC-curve are called scale inefficiencies.

[3] In the early studies of economies of scale, all banks were assumed to operate on the efficient, i.e. the lowest, AC-curve.

Figure 3.1



Liebenstein (1966) created closely related concept of X-inefficiency to represent cost inefficiencies that are due to wasteful use of inputs or managerial weaknesses. Both Farrell's technical efficiency and Liebenstein's X-efficiency seek to explain why all firms do not succeed in minimizing the costs of production. Even though technical efficiency and X-efficiency are usually thought as a same concept, they have slightly different backgrounds. Farrell's technical inefficiency raises from the firm's own actions. The efficiency problem is then technical in nature and the management should thus be able to solve it by reorganizing the production. Liebenstein's X-efficiency on its half notices that sources of X-efficiency may also lay outside of the firm. Thus, X-inefficiency raises partly from firm's own actions but also from exogenous effects of the surrounding environment. (Button and Weyman-Jones 1992).

3.3 Allocative efficiency

The concept of allocative efficiency tells whether a producer uses inputs in right proportions. Assume that input prices are fixed and producer's objective is to minimize costs. The production is then said to be allocatively efficient if costs cannot be lowered by changing the input ratios used in production. Thus, allocative efficiency tells the optimal input combinations. The marginal rate of technical substitution ($MRTS_{ij}$) must equal the ratio of the input prices, that is

$$MRTS_{ij} = \frac{f_i(x^0)}{f_j(x^0)} = \frac{w_i}{w_j} \qquad (2)$$

Allocative inefficiency results from employing inputs in the wrong proportions, which is costly. Observed costs of production are above minimum, because some inputs are used relatively too much and some too little. Since costs are not minimized, profit cannot be maximized.

The observed expenditure of the firm, $w'x^0$, can only equal the minimum cost determined by $c(y^0, w)$, if and only if the firm is both technically and allocatively efficient. If $w'x^0 > c(y^0, w)$, the difference may be due to technical inefficiency alone, allocative inefficiency alone, or some combination of the two.

Several reasons can cause a firm to be technically or allocatively inefficient. The most often mention reason is bad management, which fails to control costs and allocate resources efficiently. People doing the same work have different routines, habits and motives, which results in different amount of outputs. It is management's task to organize production so that all resources are fully utilized and that motivation of the employees is high. The second reason for technical and allocative inefficiency is that efficient production technology is at least partly unknown. The best practise technology can only be found by trying alternative technologies, which necessarily includes at least some waste or misallocation of inputs. In addition, incompleteness of input markets may lead to situation where, for example, good management may be rarely available or companies may have different possibilities to recruit efficient managers. Also uncertainty or other reasons may cause firms to imitate each other's inefficient technologies.

3.4 Scale efficiency

Firms can alter their the level of output either by changing the scale of production by varying all inputs in the same proportion or by changing the relative input proportions. Technical and allocative efficiencies measure the input utilization at certain level of output. Scale properties of production can be studied by changing all inputs proportionally. If output increases relatively more than inputs then the firm operates with increasing returns to scale. If the relative changes in inputs and output are equal the returns to scale are constant. And finally, in the case that output increases less than inputs, the returns to scale are said to be decreasing.

A combination of technical and allocative efficiency is necessary but not sufficient for profit maximization. It is not sufficient because firm could still be scale inefficient. A firm is scale efficient when its average costs are minimized and it is operating with locally constant returns to scale. In the figure 3.1 output level Q_2 represents optimal scale of production, since average costs cannot be lowered by producing less or more output. Both firms C and D are scale efficient, but only firm D's production plan is optimal for profit maximization since it is also technically efficient. A measure of scale efficiency shows how close the firm actually is to optimal scale.

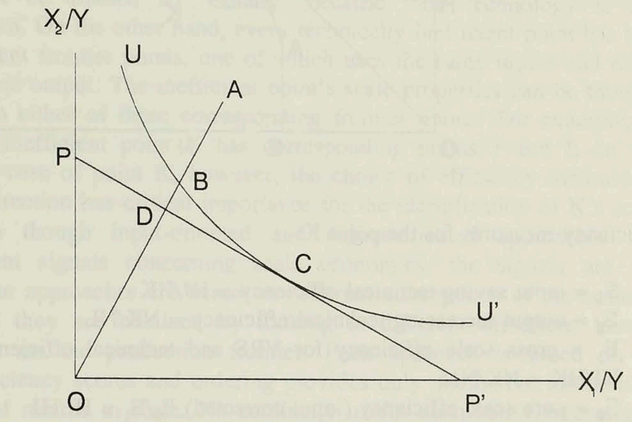
3.5 Farrell measures of efficiency

The study by Farrell (1957) created basic concepts for efficiency measurement and discussion of frontiers. Farrell defined technical and allocative efficiencies and developed calculation methods for them. His idea was to measure efficiency as a relative distance from the efficient frontier by keeping the input

proportions fixed. In his first analysis Farrell assumed that production technology is known and that returns to scale are constant.

Suppose that the firm produces output y from inputs x_1 and x_2 (only two inputs for graphical convenience). The production frontier of the firm is characterized by function $y = f(x_1, x_2)$. When returns to scale are constant, the isoquant UU' in figure 3.2 illustrates the efficient production technology for production of one unit of output ($1 = f(x_1/y, x_2/y)$).

Figure 3.2



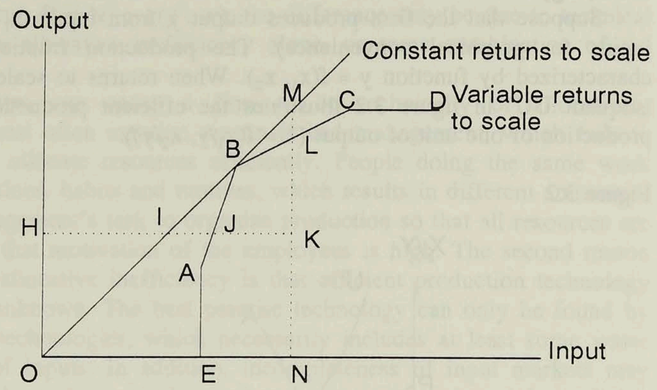
In figure 3.2 the firm produces one unit of output at point A with inputs (x_1^0, x_2^0). The point B on the isoquant UU' uses inputs in same proportions as A and also produces one unit of output. Ratio OB/OA measures the technical efficiency of the production at point A. OB/OA compares the minimum inputs required for production of one unit to the observed input usage in the firm. Thus 1 – OB/OA, the technical inefficiency, measures the proportion of inputs that could be reduced without reducing output.

The line PP' represents the ratio of input prices. The ratio OD/OB then measures allocative efficiency of the firm's input usage. The costs in point D are equal to the costs in the allocatively efficient point C, but lower than in point B. The measure of allocative inefficiency 1 – OD/OB tells the possible input savings that could be reached if the inputs were used in right proportions.

A measure for total efficiency can be obtained by adding technical and allocative efficiencies together. In figure 3.2 the total efficiency is represented by ratio OD/OA. Total inefficiency 1 – OD/OA reveals total waste of inputs, thus shows how much costs could be cut if the firm operated in the efficient point C instead of point A.

As mentioned above Farrell's efficiency measures have been generalized to cover also the variable returns to scale. In figure 3.3 efficiency measures (in a case of one input and one output) are represented relative to frontier EABCD, which exhibits variable returns to scale (VRS). The line through points O and M represents frontier of constant returns to scale (CRS).

Figure 3.3



Efficiency measures for the point K:

E_1 = input saving technical efficiency = HJ/HK
E_2 = output increasing technical efficiency = NK/NL
E_3 = gross scale efficiency for VRS and technical efficiency for CRS = HI/HK = NK/NM
E_4 = pure scale efficiency (input corrected) E_3/E_1 = HI/HJ
E_5 = pure scale efficiency (output corrected) E_3/E_2 = NL/NM

Under variable returns to scale efficiency can be measured in either input or output direction. When efficiency is measured in input direction all points on the efficient frontier EABCD fill the condition

$$E_1 = \text{Min}_\alpha \{\alpha: F(\underline{y}, \alpha\underline{x}) = 0\},$$

where $F(\cdot)$ is the transformation function. All inputs used in production are decreased in the same relative proportions (α) to reach the minimum input usage while keeping the output fixed level. The minimum input usage determines the reference technology. For example in the figure 3.3, the output level of point K could be produced with inputs of the point J on the relatively efficient frontier. The distance between the points tells the technical inefficiency for the point K.

If efficiency is measured in output direction, the points on the frontier are defined by

$$E_2 = \text{Min}_\beta \{\beta: F((1/\beta)\underline{y}, \underline{x}) = 0\}$$

In this case output is increased to maximum without changing the amount of the inputs. Thus, maximum output determines the efficient reference technology for all input combinations. Again graphical example in figure 3.3 illustrates that for instance point L is output efficient compared to point K. The values of efficiency measures E_1 and E_2 are always between zero and one.

Efficiency measure E_3 tells the overall scale efficiency in the case of VRS and technical efficiency in the case of CRS. Scale efficiency indicators E_4 and E_5 measure the distance between CRS-frontier and VRS-frontier. These measures show the pure scale efficiency since all the points on the VRS frontier are technically efficient. Naturally the scale efficiency measures can be calculated only in the case of variable returns to scale. In addition, only the technology of the points on the frontier is known with certainty, and therefore, only the frontier points' scale properties can be revealed "exactly". Scale properties of the points that are technically inefficient, i.e. lay under the frontier, cannot be studied so "exactly" because their technology is not accurately known. On the other hand, every technically inefficient point has two relatively efficient frontier points, one of which uses the same inputs and other that has the same output. The inefficient point's scale properties can be thought to be similar to either of these corresponding frontier points. For example, in the figure 3.3 inefficient point K has corresponding points J and L on the frontier. In the case of point K, however, the choice of efficiency measure or measurement direction has critical importance for the identification of K's scale property. Even though input-oriented and output-oriented approaches can provide different signals concerning scale economies, the signals are not inconsistent. The approaches are based on different conceptions of the scale of production and they are obtained by looking in different directions toward different points on the production frontier. Each approach is based on an ordering of efficiency scores and ordering provides only qualitative information on the nature of returns to scale. No cardinally useful quantitative information comparable to the economic notation of a scale elasticity or a cost elasticity is provided.[4]

[4] Førsund and Hjalmarsson (1979) propose a way to establish a single expression for scale properties, which determines the average scale property using the relationship $\varepsilon = \ln E_2/\ln E_1$. Thus, if $E_1 > E_2$ then the unit operates with increasing average returns to scale and with decreasing average returns to scale if $E_1 < E_2$.

4 Data envelopment analysis

Since the absolutely efficient production technology is unknown, the basic requirement for efficiency evaluation is construction of reference points that define the best practice production technology. Data envelopment analysis (DEA) is a non-parametric, deterministic methodology for determining the relatively efficient production frontier. DEA was originally developed by Charnes, Cooper and Rhodes (1978). They described DEA as a mathematical programming model that provides a new way for estimating extremal relations from observational data. The relation under study can be, for example, the efficient production possibility surface.

Variety of DEA-models have been developed. Here two basic models, CCR-model (after Charnes, Cooper and Rhodes, 1978) and BCC-model (after Banker, Charnes and Cooper, 1984), are introduced. These models will be used in the empirical section of this study. Other types of models will not be reviewed in this context. For further interest see e.g. Charnes et al. (1985) for additive model and Charnes et al. (1983) for multiplicative model. Seiford and Thrall (1990) discuss some recent developments in DEA.

4.1 CCR-model

Charnes, Cooper and Rhodes (CCR) (1978) introduced a measure of efficiency for each decision making unit (DMU) that is obtained as a maximum of a ratio of weighted outputs to weighted inputs. This denotes that more output produced from given inputs means more efficient production. The weights for the ratio are determined by a restriction that the similar ratios for every DMU have to be less than or equal to unity. This definition of efficiency measure allows multiple outputs and inputs without requiring preassigned weights. Multiple inputs and outputs are reduced to single 'virtual' input and single 'virtual' output by optimal weights. The efficiency measure is then a function of multipliers of the 'virtual' input-output combination. Formally the efficiency measures for DMU 0 can be calculated by solving the following mathematical programming problem:

$$\max_{u,v} h_0 = \frac{\sum_{r=1}^{s} u_r y_{r0}}{\sum_{i=1}^{m} v_i x_{i0}} \qquad (3a)$$

subject to

$$\frac{\sum_{r=1}^{s} u_r y_{rj}}{\sum_{i=1}^{m} v_i x_{ij}} \le 1 \qquad (3b)$$

where u_r, $v_i \ge 0$ (weights); r = 1,...,s (outputs); i = 1,...,m (inputs) and j = 1,...,n (DMUs).

The variables u_r and v_i are the weights to be determined by the programming problem. The x_{ij} and y_{rj} are the observed inputs and outputs of the DMU j. This problem, however, has infinite number of solutions since if (u^*,v^*) is optimal then also $(\alpha u^*, \alpha v^*)$ is optimal. The linear fractional programming problem above can be transformed into equivalent linear programming problem by selecting a representative solution (u,v) for which $\Sigma_i v_i x_{i0} = 1$.[5] Thus, denominator in the above efficiency measure h_0 is set to equal one. By adding this new restriction and reorganizing the equations, the transformed linear problem for DMU 0 can be written:

$$\max_u z_0 = \sum_{r=1}^{s} u_r y_{r0} \qquad (4a)$$

subject to

$$\sum_{i=1}^{m} m_i x_{i0} = 1 \qquad (4b)$$

$$\sum_{r=1}^{s} u_r y_{rj} - \sum_{i=1}^{m} v_i x_{ij} \le 0, \qquad j = 1,...,n \qquad (4c)$$

$$u_r \ge 0,\ v_i \ge 0\ \forall\ r,i \qquad (4d)$$

As every linear programming problem, the above problem has a corresponding dual problem, which can be written for DMU 0 as:[6]

$$\min_{\lambda}\ \theta_0 \qquad (5a)$$

subject to

[5] This transformation was introduced by Charnes and Cooper (1962).

[6] For the existence, duality and complementary slackness theorems of linear programming problems see e.g. Intriligator (1971, 77–86). A simple numerical example of solving CCR- and BCC-models is presented in the appendix 1.

$$\sum_{j=1}^{n} \lambda_j y_{rj} \geq y_{r0}, \qquad r = 1,...,s \tag{5b}$$

$$\theta_0 x_{i0} - \sum_{j=1}^{n} \lambda_j x_{ij} \geq 0, \qquad i = 1,...,m \tag{5c}$$

$$\lambda_j \geq 0 \tag{5d}$$

Both of these linear problems yield the optimal solution θ^*, which is the efficiency score for a particular DMU. The value of θ is always less than or equal to unity. The programming problems are repeated for each DMU j = 1,...,n. DMUs for which $\theta < 1$ are inefficient, while DMUs for which $\theta = 1$ lay on the frontier and are relatively efficient. The frontier consists of linear facets, that are determined by the efficient units of the data. Thus, the resulting frontier production function has no unknown parameters.

4.2 BCC-model

The linear programming problem for θ represents a solution for efficiency measure E_3 presented in section 3.5. There are no constraints for the weights λ_j other than that they are to be positive. This is the case for constant returns to scale. For variable returns to scale the sum of λs must be constrained to equal one. So, for the calculation of efficiency measure E_1 from the section 3.5, this additional constraint must be added. The resulting DEA model that exhibits variable returns to scale is called BCC-model after Banker, Charnes and Cooper (1984). The linear programming problem for Farrell input saving technical efficiency measure E_1 (variable returns to scale) for unit 0 can be written formally:

$$\text{Min}_{\lambda_0} \; E_{1_0} \tag{6a}$$

subject to

$$\sum_{j=1}^{n} \lambda_{0j} y_{rj} \geq y_{r0} \qquad (r = 1,...,s) \tag{6b}$$

$$E_{1_0} x_{i0} \geq \sum_{j=1}^{n} \lambda_{0j} x_{ij} \qquad (i = 1,...,n) \tag{6c}$$

$$\sum_{j=1}^{n} \lambda_{0j} = 1 \tag{6d}$$

$$\lambda_{0j} \geq 0 \qquad (j = 1,...,n) \tag{6e}$$

The first constraint states that output of the reference unit must be at least at the same level as the output of unit 0. The second constraint tells that the efficiency corrected input usage of the unit 0 must be greater than or the same as the input use of the reference unit. Since the correction factor is same for all types of inputs, the reduction in observed inputs is proportional. The third constraint ensures convexity and thus introduces variable returns to scale. If convexity requirement is dropped the frontier technology changes from VRS to CRS. The efficiency scores always have smaller or equal values in the case of CRS, which is also intuitively clear from the figure 3.3 above.

In the case of VRS efficiency can also be measured into output direction. In figure 3.2 this can be demonstrated by moving from the point K vertically to point L. Output increasing efficiency measure E_2 can be calculated in a similar manner as E_1 by solving another linear programme, which maximizes the amount of output from the available inputs (see Førsund 1992, 34).

Referring back to figure 3.3, the input saving technical efficiency measure E_1 with CRS as a reference technology is equal to the gross scale efficiency measure E_3 with VRS as a reference technology. Thus, after solving the linear programming problems for the measures E_1, E_2 and E_3, the measure of pure scale efficiency, E_4 or E_5, can easily be calculated from the knowledge of E_1, E_2 and E_3.

Scale inefficiencies exist if DMU's average costs are above minimum. The DMU is then be operating either with increasing or decreasing returns to scale. Scale properties of the VRS frontier points can be studied by examining the sum of weights, $\Sigma\lambda_j$, for the E_1 measure with CRS technology. If the sum of λs on the CRS frontier is less than one, then the adjusted point on the VRS frontier experiences increasing returns to scale. A sum less than one means that the best practice points determining the CRS frontier technology are scaled downwards when defining the reference points on the CRS frontier. For example in figure 3.3 point B is the only one of the data points A–D that experiences constant returns to scale, i.e. $\Sigma\lambda_j = 1$. In order to inspect the scale properties of point J, the best practice point B is scaled down to reference point I with a weight OI/OB < 1. Similarly, for the reference point M, point B must be scaled up with a weight OM/OB > 1, which indicates that VRS frontier point L experiences decreasing returns to scale (Førsund 1992). Appendix A illustrates how to examine the scale properties in the case of a simple numerical example.

4.3 Advantages and limitations of DEA

DEA provides an alternative to conventional econometric methods. Statistical analysis relies on central tendencies, whereas DEA is based on extremal

observations. In the parametric approach a single estimated regression equation is assumed to apply to each unit in the data. DEA, however, optimizes the performance of each unit separately and produces individual efficiency measures with respect to the entire set under evaluation.

The regression approach requires a priori assumption about the analytical form of the production function. This need not be easy since the underlying functional form of the truly efficient technology is usually unknown. DEA constructs the best practice production function entirely on the basis of observed data and therefore the possibility of misspecification of the production technology is minimized. The greatest advantage of the non-parametric approach is simply that the analysis is not restricted by any functional form or assumptions on the distributions of inefficiencies or random error. On the other hand, since no statistical assumptions are made, the tools of statistical inference cannot be used either. The robustness of the results has to be examined by checking the sensitivity of the analysis to different variations in the model specification, variable aggregation and the data set.

The principal disadvantage of DEA is that the frontier is sensitive to extreme observations and measurement errors. In the parametric approaches assumptions have to made about the distributions of inefficiency and random error. In the DEA approach, the basic assumption is that random errors do not exist and that all deviations from the frontier indicate inefficiency. Thus, possible measurement errors in the observations are interpreted either as excess efficiency or excess inefficiency. Because of this, much effort has recently been devoted to the development of a stochastic DEA-models (see e.g. Lovell 1993, pp. 34–35 for short overview). However, empirical evidence of these models is still quite limited.

The regression and data envelopment approaches are not necessarily competing methods, but rather the choice of the method depends on the subject matter and the quality and the amount of the available data. The estimation accuracy of either method depends greatly on the level of measurement error. Parametric approach and DEA have been compared by Banker et al. (1993). They applied corrected ordinary least squares (COLS) and BCC-model of DEA to simulated data. The results indicated that DEA is the better method of the two for most cases even with relatively high measurement errors. However, DEA's relative accuracy and comparative advantage as an estimation method enhances in cases of low measurement errors. The accuracy of COLS typically grows with increasing sample sizes, but the method generally fails to decompose deviations into efficiency and measurement error components.

The assumption about non-existing random errors in the DEA approach slightly reduces the reliability of individual efficiency rankings. There are, however, some ways to check the robustness of the efficiency rankings in general. Sensitivity of efficiency scores can be tested by deleting one or more of the original frontier observations and then re-running DEA for the remaining data. Radically changing results can be a sign of large measurement errors in the frontier observation or of some exogenous reason causing abnormal fluctuation in the observation. The level of aggregation of input and output variables may also affect the results of the DEA. The influence of the number of variables can be similarly tested by varying the input-output model and checking the correlation of the outcomes. The results can be considered robust if the conclusions of different models are not contradictory.

A general feature of the DEA in the case of VRS is that units at the both ends of the size distribution may be identified technically efficient simply for lack of other comparable units. Also, since DEA establishes a convex "lid" over the observations, the scale properties of the units are bound to occur in the order of increasing, constant and decreasing returns to scale starting from the smallest units. Because of convexity at least one unit of the data is determined to be scale efficient. Of course it is possible that only the largest unit is operating with constant returns to scale and thus the industry would appear to operate with increasing returns to scale. But if the scale efficient units are middle sized there is an identification problem whether the scale inefficiency of the technically efficient large and small units is real or whether it is due to the VRS specification and the method of enveloping the data. This problem arises especially when data is thin on very small and very large units. So, if there are e.g. only few very large units, they will probably appear as technically efficient while scale inefficiency can be substantial. Robustness of the results for the extreme size classes could be qualitatively evaluated by checking the number of units in those classes.

DEA can easily handle the case of multiple outputs and inputs, which is a feature of great relevance for banking industry studies. Even though measurement errors may affect some individual efficiency rankings in DEA, one of the safest sides of the method is that it avoids the danger of distorting the evidence by imposing a wrong parametric form and wrong distributional assumptions about inefficiencies and random errors. Since we don't have exact knowledge about the analytic form of production function, there are rational grounds for letting the data to determine the best practise technology.

4.4 Malmquist productivity index

Productivity growth in banking sector is often studied by analyzing different cost ratios obtained from accounting data (e.g. Revell 1980). These partial factor productivity (PFP) ratios can be for example such as granted loans per employee or total deposits per branch. Various possible indicators typically measure only some aspect of productivity and thus fail to recognize the multi-output nature of banking. All partial productivity studies are weakened by their inability to account for the cost of generating changes in e.g. labour productivity. For example, if a bank replaces labour with machines to carry out routine functions, it may raise labour productivity, but with no particular change in overall costs.

Total factor productivity (TFP) is a generalisation of the PFP-ratios. It extends the concept of PFP by incorporating multiple outputs and multiple inputs in to a single productivity ratio. The central issue in TFP measurement is the choice of the method that is used to estimate the weights used to combine inputs and outputs. The advantage of TFP over PFP measures is that it enables consistent productivity comparisons to be made across the range of banks' outputs and inputs, whereas a priori there is nothing to guarantee that the corresponding PFP ratios will give a consistent picture of productivity performance. One PFP-ratio may show very high productivity but another very low and therefore comparison of banks is complex. However, calculation of

TFP over time is difficult because proportions of factor inputs do not remain constant over time, and their contribution to output is difficult to unravel (Colwell and Davies, 1992). Previous studies of total productivity growth in banking are mostly based on the estimation of cost functions (e.g. Kim and Weiss, 1989). This approach is appropriate for identifying the average practice productivity growth, but not for identifying productivity growth at the best practice banks.

A change in total productivity can be caused either by technological progress or by a change in relative technical efficiency of the production unit. For identification of productivity growth between two time periods the Malmquist productivity index can be used. The original index developed by Malmquist (1953) measured the quantity of consumption that consumer should achieve in certain year to get the same utility level as in the year before. This proportional scaling factor or quantity index was a ratio of two distance functions in different time periods. Caves et al. (1982) developed Malmquist idea into a productivity index. They utilized Shepard's concept of distance functions when defining proportional scaling. They did not, however, notice the direct link with Farrell (1957). This connection was pointed out by Färe et al. (1985).

Caves et al. separate input-based and output-based productivity indices, which is in line with Farrell's input saving and output increasing efficiency measures. The productivity index is based on comparisons between two different points in time, only quantities are involved and at least either one of the technologies has to be known. The idea of what Caves et al. named the Malmquist unit 1 input based productivity index is to find minimal proportional scaling of inputs for unit 2 such that its scaled input vector and its observed output vector are just on the production surface of unit 1. Caves et al. assumed that units were operating efficiently.

Färe et al. (1989) applied Malmquist index approach to nonparametric framework and extended it to allow inefficient observations by replacing the assumed efficient technology with relatively efficient frontier technology. In both periods there can then be observations under the frontier i.e. observations from relatively inefficient units. Each period's observations must be adjusted to the applied frontier technology. The required frontier technologies can be constructed with DEA. This approach has been applied to Norwegian banking by Berg et al. (1992) and to comparing banking productivities in the Nordic countries by Berg et al. (1993).

The definitions of Farrell's efficiency measures can be used directly in the definition of the Malmquist productivity index in the nonparametric framework. Malmquist input based productivity index M_i with period's i frontier technology $F_i(\cdot) = 0$ as reference is

$$M_i = \frac{E_{i2}}{E_{i1}} = \frac{\mathrm{Min}_{\alpha_{i2}}[\alpha_{i2}: F_i(y_2, \alpha_{i2}x_2) \le 0]}{\mathrm{Min}_{\alpha_{i1}}[\alpha_{i1}: F_i(y_1, \alpha_{i1}x_1) \le 0]}, \qquad i = 1,2 \tag{7}$$

Either first or second period frontier technology can be used as reference (i = 1 or 2). The numerator tells the relative adjustment of observed input usage in the period 2 that would be required to shift the observation to the frontier i, while output is kept constant. The denominator tells, in the similar manner, the

relative adjustment of observed input usage in the period 1 that would shift the observation to the frontier i. The measures may be greater than one in the cases when the first year's observation is compared to second year's frontier or when the second year's observation is compared to first year's frontier. If $M_i > 1$, then production in the period 2 is more productive than production in period 1. The definition for output based Malmquist index can be obtain similarly from Farrell's output increasing efficiency measure (see Førsund 1990).

The Malmquist productivity index can decomposed into the shift of the frontier and the change in efficiency relative to frontier as in Nishimizu and Page (1982):

$$M_i = CU * FR_i, \qquad i = 1,2 \tag{8}$$

Where CU is the catching up component that measures the change in relative efficiency and FR_i is the productivity change resulting from the frontier's movement.

$$CU = \frac{E_{22}}{E_{11}}, \; FR_i = \frac{E_{1j}}{E_{2j}}, \qquad i,j = 1,2, \; i \neq j. \tag{9}$$

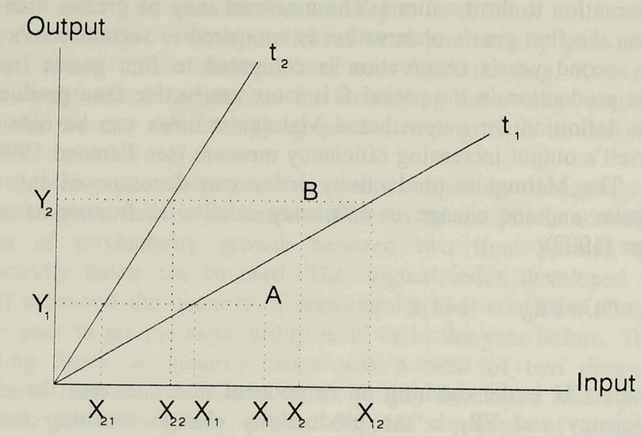
CU shows the relative movement towards the frontier since it is a ratio of efficiency measures relative to each period's own frontier. FR_i measures the distance between the frontiers by comparing the same observation to both frontiers.

Figure 3.4 presents graphically how productivity measures are obtained in the simple case of one input and one output. The graph illustrates two frontiers (constant returns to scale) t_1 and t_2 at two time points. Points A (x_1, y_1) and B (x_2, y_2) show respectively observed inputs and outputs of the production unit at these two periods. In the same manner as in figure 3.3, the ratios x_{11}/x_1 and x_{22}/x_2 represent the technical efficiency scores E_{11} and E_{22} for points A and B relative to the frontiers t_1 and t_2. Ratios x_{21}/x_1 and x_{12}/x_2 give efficiency in terms of the frontier of the other period and thus represent the efficiency scores E_{21} and E_{12}. The Malmquist productivity index can then be written as:

$$M_1(t_1, t_2) = \frac{E_{12}}{E_{11}} = \frac{\dfrac{x_{12}}{x_2}}{\dfrac{x_{11}}{x_1}} = \frac{\dfrac{y_2}{x_2}}{\dfrac{y_1}{x_1}} \tag{10}$$

The last ratio is derived utilising the relationship $x_{12}/x_{11} = y_2/y_1$. The Malmquist index is, in the case of constant returns to scale, simply a ratio of productivities in the two periods and thus tells the change in productivity between the two periods. Thus in this one input case the Malmquist index is identical to standard factor productivity index.

Figure 3.4



Furthermore, frontier and catching up productivity indices can, from figure 3.4, be written as:

$$FR(t_1,t_2)=\frac{E_{12}}{E_{22}}=\frac{\frac{x_{12}}{x_2}}{\frac{x_{22}}{x_2}}=\frac{x_{12}}{x_{22}},\quad CU(t_1,t_2)=\frac{E_{22}}{E_{11}}=\frac{\frac{x_{22}}{x_2}}{\frac{x_{11}}{x_1}} \tag{11}$$

In the multiple output and input setting problems may arise if the input and output structures have changed much during the period. This may lead to problems with respect to consistent comparison of the unit's observations. If the relatively efficient frontiers for different technologies intersect, it is even possible that Malmquist indices relative to these reference technologies give opposite results. Critique has been presented about that the units being compared in the efficiency analysis do not necessarily produce the same level of output. If constant returns to scale technology is assumed an efficient small bank may be a reference unit to a very large bank. If inefficiency is assumed to be independent of output level, then the relative efficiency component of the malmquist index would also measure the relative productivity of the units (see Mester 1993).

5 Measures of bank production

A key question in bank efficiency studies has been how to measure bank production. There are two major approaches to this problem, the so called production approach, which concentrates on the bank's production of services to the public, and the so called intermediation approach, which concentrates on the role of banks as financial intermediators.

5.1 Production approach

The production approach suggests that the most important function of banks is the production of services. These services consist of various kinds of payment services usually related to deposit accounts and of different savings opportunities offered by a bank. Another class of services according to production approach is the loan supply to public and firms. Banks monitor and evaluate the risks of investments, take care of collecting payments and interest of invested funds, offer the benefits of efficiently diversified portfolio also for small savers and take care of payment processing. Banks also offer such services as custody of securities, dealings with stocks and other securities, safe-keeping, and various kinds of counselling services. This kind of service production uses real resources as inputs. These are traditionally measured by the use of labour, materials and capital invested in buildings and machinery such as data processing machines, atm's and electronic networks between banks. The production approach underlines the importance of physical services and therefore measures the costs of banking by operating costs. Outputs are typically represented by either the number of different types of loans and deposit accounts or by the money values of these two. Output vector may also include either number transactions or value of provisions from various services such as payment processing or intermediation of stocks or other securities.

Neither of alternative proxies of bank production is without problems but especially the number of accounts approach has a few obvious weaknesses. If number of accounts measure is correct, the physical bank production process should have one to one correspondence with the number of balance sheet items. This is highly questionable, because the distributions of these balance sheet items and the underlying transactions are not symmetric. For example most of the deposit accounts are small in money value, but large number of transactions are processed through these accounts. On the other hand, small percentage of deposit accounts are big in money value and these accounts are stable. So it is not without problems to infer how the costs of a bank are related to the number of balance sheet items. There is also large number of so called "dead accounts", which are not used and do not have deposited funds. Because having a bank account does not cost anything, there are a lot of these kind of accounts. According to the banking law bank can close this kind of an account after it has not been used for ten years. When the share of these accounts varies between banks and can be very high, it is clear, that the mere number of different types of deposit accounts is not very informative. So, using the money values of

deposit accounts, loans and other relevant balance sheet items should be preferred to the alternative of the number accounts in the production approach.

5.2 Intermediation approach

The intermediation approach emphasizes the important function of banks as an financial intermediaries. In this case the costs also must include interest costs in addition to operating costs. According to this approach the magnitude of intermediated funds in different types of loans and securities measure the production of a bank. Money values of relevant balance sheet items are employed as proxies for these products of banks.

Production and intermediation approaches stress different aspects of banking and should be seen as completing each other. If we want to reach conclusions about the efficiency of the physical service production process, we must follow the production approach. The ongoing changes in banking, especially in Europe and United States, highlight the importance of this approach. After the deregulation banks are faced with new competitive environment where cost efficiency of physical production process plays more important role than before. Therefore it is meaningful to learn about the determinants of efficient service production. On the other hand, if the main interest is to study market structure of the banking industry, we must examine how the total costs of banks are formed. The operating costs do not give enough information with this respect, because when banks get bigger, they usually change the structure of their financing towards market funding. In this process the share of operating costs decrease, but in the same time the share of interest costs typically increases. So, using only operating costs to study market structure increases the possibility of biased conclusions that e.g. significant economies of scale exist in banking industry.

6 Empirical analysis

6.1 Data

Input and output definitions in this study follow the production approach. It is appropriate for studying productive efficiency since it concerns just operating costs of banking. Interest costs are not included because the efficiency of physical service production depends on the use of real resources. Services cover all types of bank activities from loan and deposit related services to broker services in securities markets.

Data used in this study consists of accounting data for Finnish local banks during the time span 1985–1990. The data is based on the banking statistics published by Central Statistical Office of Finland and it includes both cooperative and savings banks. Total number of local banks in 1985 was 624 decreasing through mergers to 509 in 1990. Most of the mergers took place in savings banks.

Use of the panel data in efficiency studies is worthwhile especially if it is assumed that inefficiencies are stable and random fluctuations tend to average out over time (see Schmidt and Sickles 1984, Berger 1992). For this assumption to hold it is crucial that random errors are not systematic. In data envelopment analysis use of the panel data helps to control for the effects of random variation, which is assumed not to exists in single cross-section studies. With pooled data we are able, at least to some extent, to avoid interpreting random errors as inefficiency, which cannot be avoided when DEA is applied to a single cross-section data. The pooled data was constructed by aggregating those banks that had merged during 1985–1990. Thus, the pooled data represents the structure of local banks in 1990. It includes 471 real observations, i.e. banks that had not been a part of any merger, and 38 observations that were constructed by aggregation.

The variables chosen for the input vector are:

1) Labour
2) Operating costs
3) Machinery and equipment

Labour is measured by the number of personnel. Operating costs are measured by the item 'other costs' from the income statement. It includes expenses from the use of ADP, telephone, postage and office supplies, rents and leases, real estate and marketing expenses, and also other miscellaneous expenses. The book value of machinery and equipment from the balance sheet indicates the use of capital. Building capital is ignored from the input vector, because satisfactory indicator was not available in the data set. This, however, should not be a problem, since the expenses from using the buildings are included in the operating costs.

For the measurement of bank's output two alternative output vectors are defined. The first one measures loans and deposits by their money value and the second one by the number of accounts. Output vector tries to take into account all the basic functions of banking. The traditional outputs in the

production approach are loans and deposits. Here, also other type of services are represented with a variable 'other earnings' and availability of the banking services is represented with number of branches.[7] The variables chosen for the output vector are:

1) Short term loans to non-banks
2) Long term loans to non-banks
3) Cheque accounts by the public
4) Deposits by the public
5) Number of branches
6) Other earnings

Short term loans include overdrafts and bills. Long term loans contain ordinary loans and loans granted from state funds. The deposits are divided into two groups, cheque accounts and deposits by the public. The values of loans and deposits are obtained from the year-end balance sheets. The other earnings include commissions and fees from payment processing and bank guarantees, earnings from foreign exchange and securities dealings and earnings from real estate.

None of the variables specifically measures the amount of produced payment services, which traditionally have demanded a lot of resources in Finnish banking system. However, the category of other earnings includes fees from those services that are priced explicitly. If it is furthermore assumed that each local bank produces relatively the same amount of free payment services, then the production of these services is to some extent implicitly measured by the amount of deposits and by the number of branches.

When activities are measured by the number of accounts the output vector consists of only three variables; number of loans, number of deposits and number of branches. In this case the total number of loans includes the number of bills and ordinary loans. The number of deposits contains only the number of deposits by the public and does not include the number of cheque accounts. The number accounts output vector is carried along merely in interest to be able to perform some comparison and checking of the results. However, the money value output vector is considered to be the primary measure for bank's activities.

As discussed in section 4.3 robustness of the DEA results depends on the level of measurement errors in the data. The chosen data is obtained from the official statistics and thus is the most reliable data that is publicly available. Most of the used variables, such as outstanding loans and deposits by the public, do not include much risk of being too sensitive to random errors. Number of employees and number of branches are also accurately measured. Machinery and equipment is the book value from the balance sheet and it may therefore include some valuation variation across the banks. The only variable that could include large random variation is other earnings. Some speculative earnings from e.g. from sales of properties may cause bias. However, the risk of

[7] In some studies number of branches has been identified as an input. However, it would not give much new information if it were included in the input vector in the contex of this study. On the contrary, in the output vector number of branches gives information that is not measured by other chosen output variables.

that feels smaller than the loss of leaving the whole variable out of the output vector.

Appendix 2 presents data for average bank for each cross-section through years 1985–1990. Averages are calculated for all local banks and for cooperative and savings banks separately. Loans, deposits, and other earnings were deflated by GDP-deflator, other costs and machinery and equipment were deflated by relevant indices calculated by Central Statistical Office of Finland (see appendix 2). An average savings bank appears to be larger than an average cooperative bank. Also, average savings bank has grown faster than average cooperative bank. One reason for this is that more mergers took place in savings banks and the other is that the credit expansion was faster in savings banks.

6.2 Efficiency results

Data envelopment analysis was performed for the data set described above. Input saving and output increasing technical and scale efficiency measures were calculated for six cross-sections and for the pooled data spanning years 1985–1990. The two alternative output definitions were employed for each year and efficiency measure.

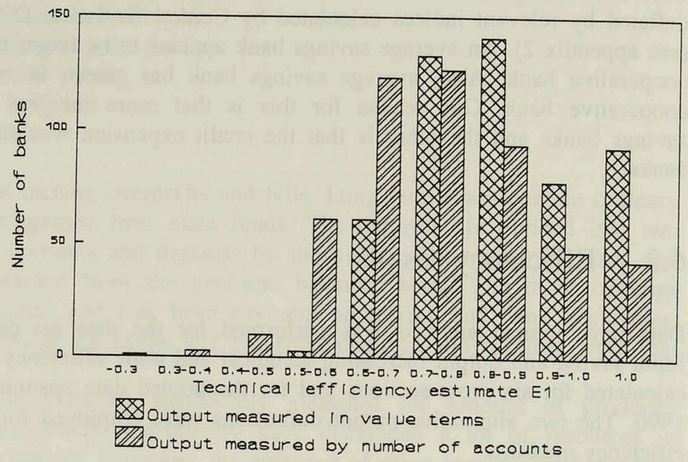
6.2.1 Differences in results due to the alternative output definitions

Given the amount of inputs an individual bank is determined to be efficient if the money value of its loans, deposits and other earnings is as large as any linear combination of other banks' output variables. The bank is inefficient if it is possible to construct a reference bank with equal amount of inputs and output as a linear combination of other banks' outputs so that this reference bank produces more output measured in value terms. In the case that bank's activities are measured by number of accounts the determination of efficient banks is analogous.

The distributions of efficiency scores for alternative output definitions are almost similar. Figures 6.1 and 6.2 show the frequency distributions of technical and scale efficiency estimates for the pooled data set in the case of variable returns to scale. The outcomes for both output definitions support the common result of other banking efficiency studies that the level of technical inefficiency is greater than the level of scale inefficiency. The average level of technical and scale efficiency is little lower when output is measured by the number of accounts. Average technical inefficiency is 15.9 % when output is measured in value terms and 24.8 % in the number of accounts case. Modes of the alternative scores are in different classes and the distribution of the number of accounts scores is more skewed to the left. This can be clearly noticed when cooperative and savings banks are examined separately (see appendix 3). For example, money value approach doesn't place any savings banks into to the four lowest efficiency classes. The distributions of the cooperative and savings banks are similar for both output definitions. Thus, according to the pooled

data, there are not much technical efficiency differences between the cooperative and savings banks.

Figure 6.1 **Frequency distribution of technical efficiency scores pooled data 1985–1990**



For the scale efficiency measures averages of the money value and number of accounts approaches are 7.3 % and 9.6 %. Distributions of the two are highly similar (figure 6.2). However, distributions differ slightly when savings banks are examined separately (see appendix 3). Money value scores show little higher frequencies in the two lowest classes. Furthermore, regardless of the output definition, scale efficiency rankings of the savings banks are on average lower than rankings of the cooperative banks.

Part of the difference in the inefficiency levels between the alternative output definitions can be explained by the different dimensions of specified output vectors. Money value output vector has six variables, while the number of accounts vector includes only three output variables. When the money value output vector is aggregated to contain only total loans and total deposits (short term and long term loans and deposits are combined so that each is represented by one variable and other earnings is left out), the average technical inefficiency raises from 15.9 % to 24.1 % so that differences between value and number measures seem to vanish. The average scale inefficiency, on the other hand, falls slightly after aggregation from 7.3 % to 6.8 %, and thus, the difference in the results of the alternative output definitions grows marginally.

Figure 6.2 **Frequency distribution of scale efficiency scores. Pooled data 1985–1990**

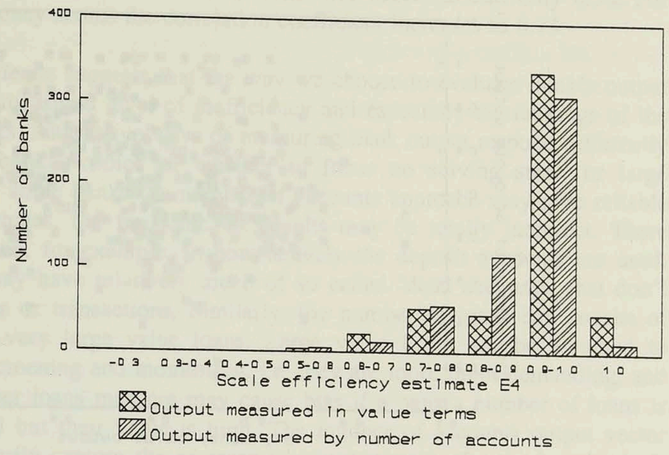


Figure 6.3 presents the correlation diagram of the two alternative technical efficiency scores. Spearman's rank correlation between the scores of individual banks with different output definition is only 0.32. Some individual banks display extremely large differences between efficiency estimates. For example, the bank in the lower right hand corner of the figure 6.3 has been ranked fully efficient with the money value output identification, but one of the most inefficient with number of accounts output identification. The difference can be explained by the fact that the bank in question has four times larger average loan size and three times larger average deposit size than a typical bank. It is apparent that this specific bank is concentrating on different type of customers than the average bank and obviously, in this case, the number of accounts output definition fails to evaluate efficiency correctly.

Also for the scale efficiency scores there are large individual differences between the two output vectors (figure 6.4). Rank correlation of the two alternative scores is 0.46. Most of the rankings are in the class of 0.9–1.0 and number of accounts output vector clearly gives lower efficiency results on average.

Figure 6.3 **Correlation diagram of E1 scores with different output vectors. Pooled data 1985–1990.**

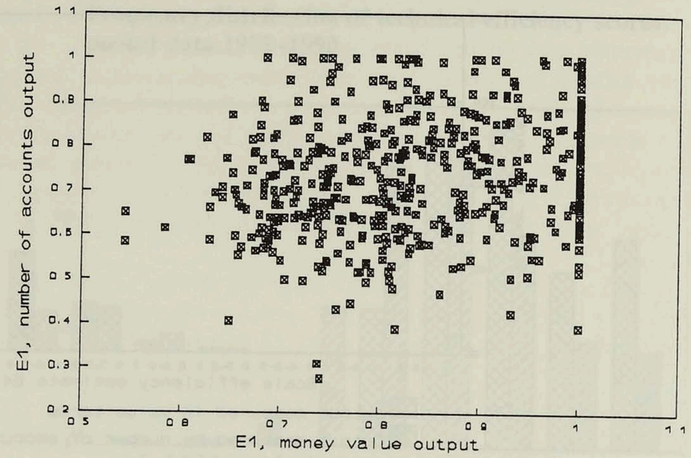
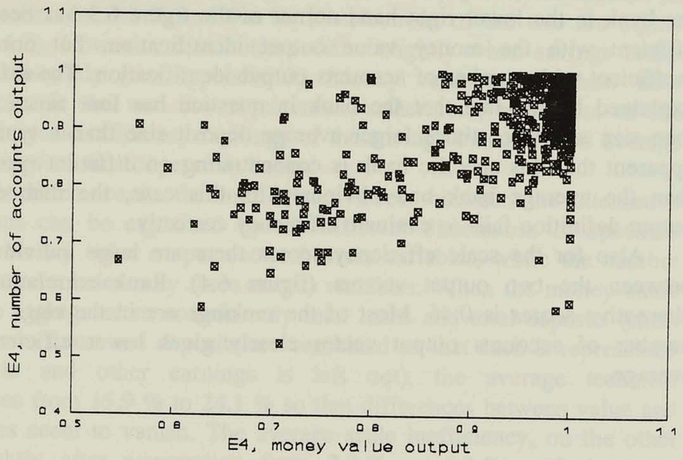


Figure 6.4 **Correlation diagram of E4 scores with different output vectors. Pooled data 1985–1990**



When the money value output vector is reduced to contain only total loans and total deposits, i.e. the comparable variables as in number of accounts vector, rank correlation between the technical efficiency scores is still only 0.56. For the scale efficiency scores the correlation coefficient increases to 0.73.

The above evidence suggests that the way we choose to evaluate bank's output influences the measured level of inefficiency and especially the rankings of the individual banks. Alternative ways of measuring bank output respond differently to individual bank's choice of whether to focus on serving small or large customers. For some banks the number of accounts approach may give reliable efficiency estimates, but for some the results may be totally incorrect. There may be variation, for example, in how actively the deposit accounts are used. Some banks may have relatively much of so called 'dead accounts' that don't have any funds or transactions. Similarly, the number of loans can consist of very small or very large value loans. Large value loans can be assumed to require more screening and monitoring, i.e. they are more input demanding, and therefore number loans measure may cause bias if a bank's number of loans is relatively small but their value is high. The number of accounts output vector doesn't necessarily capture the representative output mix of a modern bank. It should include also some indicators of other than deposit and loan related services. This type of transaction data is, however, rarely available.

The purpose of this study is not to place local banks into specific efficiency order, but rather to try to gain some insight of the common tendencies in the whole local banking sector, between the two bank groups and between the different size classes of banks. Therefore, it is not too crucial if some individual banks have bias in the efficiency score. However, robustness of the results must be checked so that there are no outliers among the frontier banks. In the following sections the money value output vector with six variables is preferred to number of accounts approach, since this output vector is assumed to represent bank's output mix much more reliably. The differences are presented only when the alternative definitions lead to exceptionally different results thus calling for qualifications in the conclusions.

6.2.2 Technical efficiency

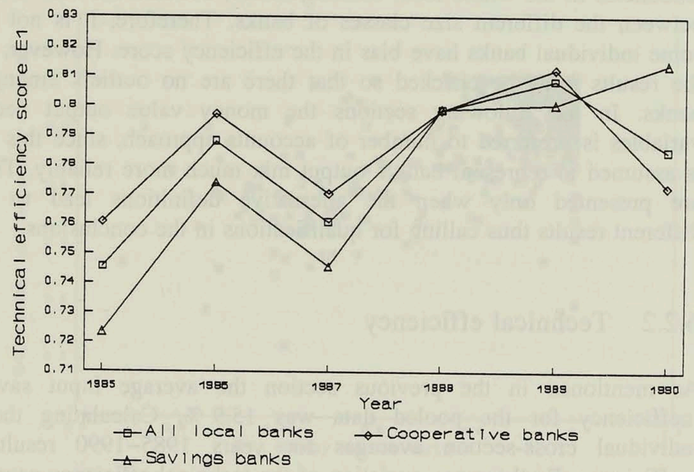
As mentioned in the previous section the average input saving technical inefficiency for the pooled data was 15.9 %. Calculating the average of individual cross-section averages for years 1985–1990 results 21.9 % of inefficiency. Furthermore, variation of the technical efficiency scores is lower in the pooled data results than in any single cross-section. This suggests that effect of random errors might be smaller in the pooled data. The errors can be assumed to be unsystematic so that their effect diminishes in time. The results of the pooled data can hence be considered to be more reliable than the results of the individual cross-section data. The single cross-sections are however worth examining for the interest of how the average or structural efficiency has changed in time.

The pooled data did not show significant differences between savings and cooperative banks' average technical efficiency scores (see appendix 3).

However, when analysis is performed for each cross-section separately some differences can be identified. Figure 6.5 shows the development of technical efficiency scores through 1985–1990. From 1985 to 1987 cooperative banks were operating on average more efficiently than savings banks. For the years 1988–1989 differences between the groups are less notable. In 1990 savings banks appear to be on average more efficient than cooperative banks. Thus, savings banks seem to have increased their relative technical efficiency towards the end of 1980's.

At this point it must be stressed that the output vectors applied in these cross-sections do not measure the quality of loan evaluations or creditworthiness of customers. Lack of risk indicator may possibly generate measurement errors for some of the banks that have expanded very rapidly. This is a potential explanation for the increase of savings banks' average technical efficiency in the last years of 1980's. Loan losses could be used to adjust loan quality indicator into the output vector. However, the problem is that loan losses arise with a lead resulting from loans given in earlier years. Length of the lead is hard to determine and thus it is difficult to specify a quality adjusted output vector for a single cross-section. In addition, Finnish local banks did not experience much loan losses until 1991 and therefore whether the losses in the years 1985–1990 are included or not doesn't influence the cross-section results.

Figure 6.5 **Average technical efficiencies 1985–1990**



Robustness of the pooled data results was checked against loan quality or riskiness by examining the correlation between occurred loan losses and efficiency levels. The loan losses of the year 1991 were chosen since it is reasonable to assume that most of the loans arising the losses were granted during 1985–1990. The result was that the technical efficiency scores obtained from the pooled data did not correlate at all with realised loan losses of 1991. This means that rapid credit expansion with poorly evaluated loans does not on average appear to influence bank's efficiency score, at least not directly. Thus,

there should not be banks that have over-ranked efficiency score due to production of bad loans. However, the same might not be the true for the single cross-section results, and therefore they must be interpreted more carefully especially in the case of individual banks.

When technical efficiency is evaluated keeping inputs fixed and scaling the outputs (i.e. output increasing technical efficiency measure E_2 is used), the average degrees of inefficiency follow similar levels as in the case of input saving technical efficiency measure E_1. Correlation of the two measures for individual banks in the pooled data results was very high (0.99).

6.2.3 The effect of technology assumption on technical efficiency

The level of bank scale economies is an empirical question, where widely differing results have been offered. The general conclusion of the literature, however, is that average cost curve is U-shaped but relatively flat. Economies of scale have been found at small banks and diseconomies at the largest banks. In the analysis above it has been assumed that banks operate with variable returns to scale (VRS). If VRS technology assumption is replaced with constant returns to scale (CRS) technology, the average level of technical inefficiency increases from 15.9 % to 22.3 %.

Figure 6.6 **Average technical efficiency vs. total loans. Pooled data 1985–1990.**

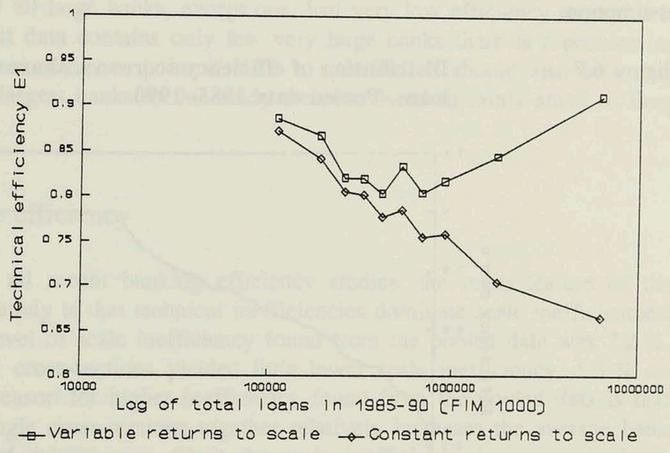


Figure 6.6 shows average efficiencies in each size decile of the pooled data. Large banks tend to show better efficiency scores than the average size bank in the case of VRS. Same kind of tendency can also be noticed among the smallest banks. Similar trends were found also from single year cross-sections. In the largest and smallest size deciles variation in size is bigger than in middle size classes. Therefore higher average efficiency scores in the large and small deciles may result from the lack of equal size reference banks, rather than being

a sign of better performance by very large or very small banks. Thus, the cause of these higher efficiency scores in both cases may rise, at least to some extent, from the methodology applied, which is used to construct the relatively efficient frontier. The differences between the size deciles should therefore not be interpreted too tightly. It would be more reasonable to conclude that there are some technical efficiency differences in favour of the largest and the smallest banks, but the differences cannot be considered to be very significant.

When the CRS-frontier is determined no convexity requirement is needed for the weight vector. This means that the frontier may be totally determined by the small banks. In other words, a large bank may have a reference unit that is a linear combination of very small banks. Here, the shape of the CRS-frontier is determined mainly by the small banks. The smallest decile of banks is found technically the most efficient and the largest decile of banks is relatively most inefficient. The difference is quite large. If the CRS assumption in local banking is correct, the smallest banks would seem to have a clear cost efficiency advantage compared to the largest local banks. On the other hand, if VRS assumption is correct, the growing deviation between the averages of the two frontier technologies towards the larger banks reflects greater scale inefficiency in the large banks. This will be examined in more detail below.

In the figures 6.7 and 6.8 the efficiency scores estimated from the pooled data are plotted against the cumulative share of total loans. When VRS technology is specified 37 % of loans are produced in technically efficient banks. Changing the technology to CRS drops the percent of efficiently produced loans to only 6 %. Thus, there are potential technical efficiency gains to be achieved in Finnish local banking regardless of the technology assumption.

Figure 6.7 **Distribution of efficiency scores vs. share of total loans. Pooled data 1985–1990.**

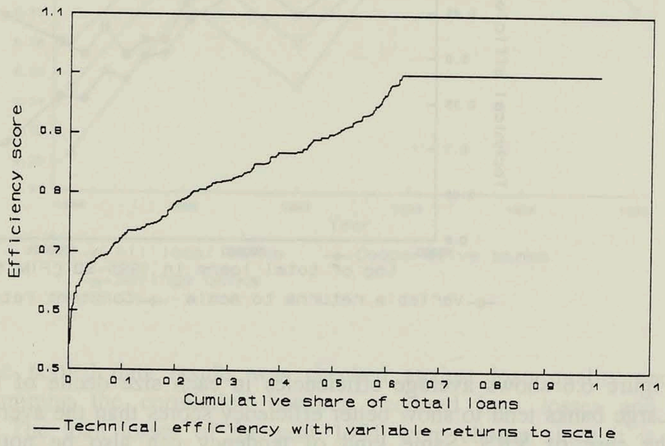
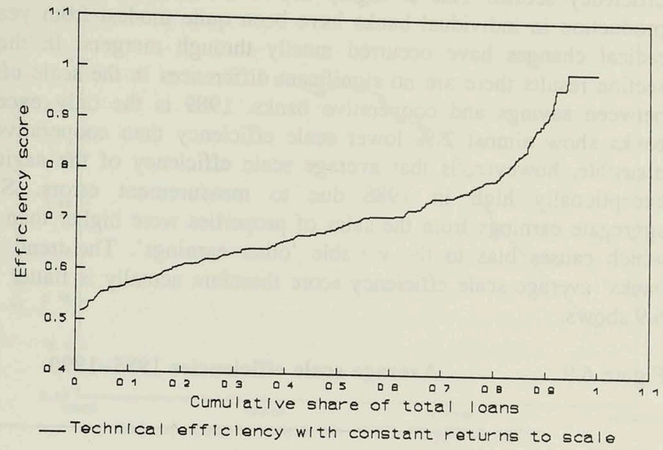


Figure 6.8 **Distribution of efficiency scores vs. share of total loans. Pooled data 1985–1990.**



In the study by Berg et al. (1993) similar analysis was performed on Finnish banks.[8] The result in the case of VRS was that about 70 % of total loans were being produced by fully efficient banks. Reason was that four largest commercial banks were evaluated to be 100 % efficient. In the case of CRS the results changed dramatically. Only 1.7 % of total loans were being produced efficiently and all large banks, except one, had very low efficiency scores. It is obvious, that if data contains only few very large banks there is a problem of constructing real reference point if VRS is specified. It should also be noted, that even the largest banks in this study are considered as rather small in Berg et al. (1993).

6.2.4 Scale efficiency

As in almost all recent banking efficiency studies, the main feature of the results in this study is that technical inefficiencies dominate scale inefficiencies. The average level of scale inefficiency found from the pooled data was 7.2 %. The individual cross-sections yielded little lower scale inefficiency, 5.3 % on average. The reason for higher inefficiency found from the pooled data is that pooling the single cross-sections together relatively increases the average bank size because of the mergers. Since the scale inefficiency is greater in large banks it is to be expected that the pooled data yields greater scale inefficiency than single cross-sections.

[8] Berg et al. (1993) used DEA to compare banking efficiency in the Nordic countries. They applied a data set for the year 1990, which included commercial banks in addition to local banks. The input and output definitions follow the production approach, but differ in some details with this study.

Development of scale efficiency from 1985 to 1990 is presented in figure 6.9. Variation in scale efficiency scores is much smaller than in technical efficiency scores. This is highly intuitive, since the changes in the scale of production in individual banks have been quite modest from year to year. The radical changes have occurred mostly through mergers. In the single cross-section results there are no significant differences in the scale efficiency scores between savings and cooperative banks. 1989 is the only exception; savings banks show almost 2 % lower scale efficiency than cooperative banks. More plausible, however, is that average scale efficiency of the savings banks was exceptionally high in 1988 due to measurement errors. Savings banks' aggregate earnings from the sales of properties were higher than usual in 1988, which causes bias to the variable 'other earnings'. The trend of the savings banks' average scale efficiency score therefore actually is flatter than the figure 6.9 shows.

Figure 6.9 **Average scale efficiencies 1985–1990**

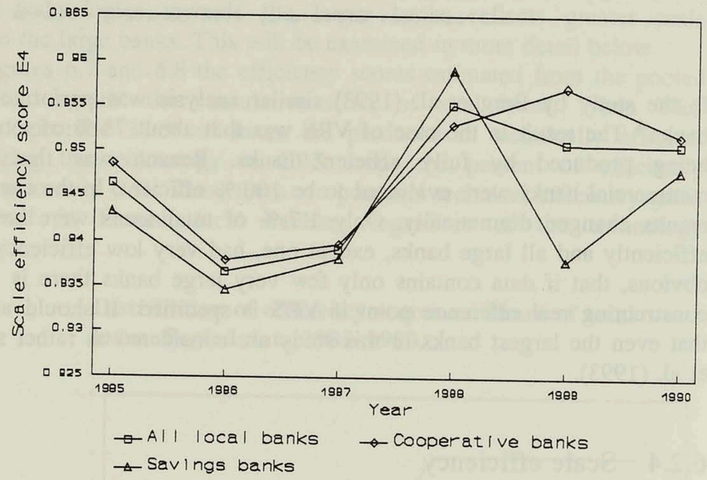
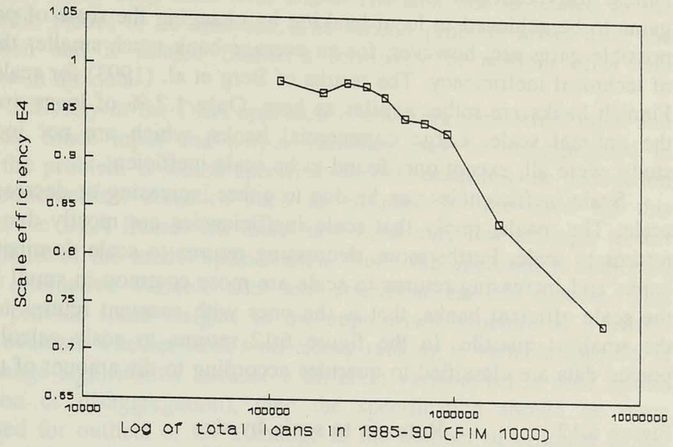
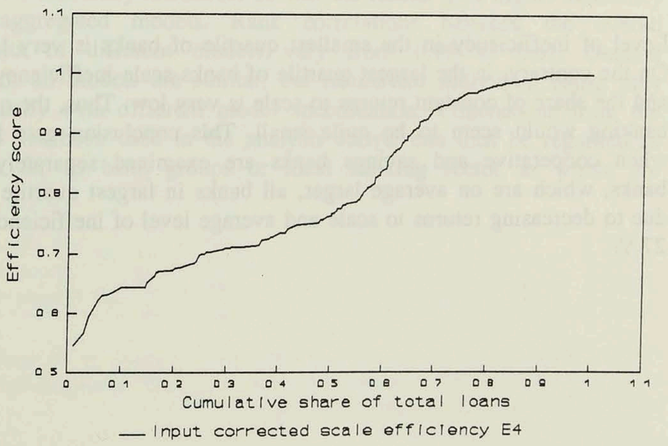


Figure 6.10 **Average scale efficiency vs. total loans. Pooled data 1985–1990**



Unlike in the case of technical inefficiency, the degree of scale inefficiency is significantly higher in the large banks. In the figure 6.10 average scale efficiency score of each size decile is presented. Scale efficiency is much lower in larger banks. Furthermore, variation of the scale efficiency scores in the small and middle sized local banks is modest, but grows for the larger banks. The results of the pooled data also suggest that average scale efficiency is 4 % higher in cooperative banks than in savings banks. As mentioned before, savings banks are on average larger than cooperative banks. Average size difference grows from its actual level when the data is pooled, since most of the mergers have taken place in savings banks. Even up to 90 % of the banks in the smallest size decile are cooperative banks and over 60 % of the largest decile are savings banks.

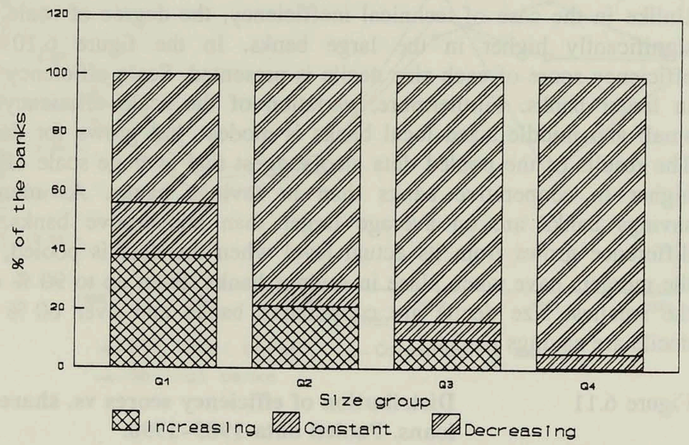
Figure 6.11 **Distribution of efficiency scores vs. share of total loans. Pooled data 1985–1990.**



Only about 10 % of the loans in pooled data were produced by fully scale efficient banks. However, about 30 % of the loans were produced by banks almost fully efficient (figure 6.11). Thus, there are some potential efficiency gains to be achieved in local banking by changing the scale of production. The possible gains are, however, for an average bank much smaller than in the case of technical inefficiency. The results of Berg et al. (1993) for scale efficiency of Finnish banks are rather similar as here. Only 1.7 % of loans are produced on the optimal scale. Large commercial banks, which are not included in this study, were all, except one, found to be scale inefficient.

Scale inefficiencies can be due to either increasing or decreasing returns to scale. The results imply that scale inefficiencies are mostly due to decreasing returns to scale. Furthermore, decreasing returns to scale dominate in the large banks and increasing returns to scale are more common in small banks. Half of the scale efficient banks, that is the ones with constant returns to scale, are in the smallest quartile. In the figure 6.12 returns to scale calculated from the pooled data are classified to quartiles according to the amount of total loans.

Figure 6.12 **Returns to scale by size**
Pooled data 1985–1990



Level of inefficiency in the smallest quartile of banks is very low, only 2 %. On the contrary, in the largest quartile of banks scale inefficiency is even 19 % and the share of constant returns to scale is very low. Thus, the optimal scale in banking would seem to be quite small. This conclusion gets further support when cooperative and savings banks are examined separately. For savings banks, which are on average larger, all banks in largest quartile are inefficient due to decreasing returns to scale and average level of inefficiency is as high as 27 %.

6.2.5 Robustness of the results

Two kinds of problems may arise with DEA. The first one is related to the choice of specific model to be analysed. The second problem arises after the model is chosen and is related frontier's sensitivity to random error and possible outliers in the data.

The main difficulty in the DEA approach is choosing the correct input and output variables. Since input and output variables may be disaggregated in several ways, the problem is which specification best describes the production relation. In econometrics, statistics such as adjusted R^2 are used for this purpose, but in the DEA framework there have been only few attempts to test the goodness of fit of the model specification. Here, the approach is subjected to variety of alternative variable sets and specifications i.e. models. This procedure can provide some insights to the appropriate structure of a bank's efficiency and evaluate whether bank's efficiency ranking is variable specific. If conclusions change significantly because a different variable set is used (either through addition or disaggregation), then the specification should be more closely examined for outliers or for a change in the number of efficient banks. In other words, for a model to be considered robust, it must be shown that reasonable changes in the list of variables cannot alter the conclusions fundamentally.

In section 6.2.1 the money value output specification was chosen as preferred output approach. However, the number of variables in the output vector and the level of aggregation seems to influence the degree of average efficiency. To see how the average level of efficiency changes with slightly different output vectors, DEA was performed for eight different money value output specifications. Loans and deposits were present in each model either in an aggregated or disaggregated form. Variables 'other earnings' and 'number of branches' were included in some and left out from some models.

Table 6.1 shows the average results from different model specifications for the pooled data. All models give similar results for relationship between technical and scale efficiency levels. Technical efficiency decreases when the number of variables is reduced (or when data is aggregated). In the case of scale efficiency aggregation does not change the results significantly but omitting variables seems to increase efficiency a little. For overall efficiency the effect of technical efficiency influences so that the results give higher efficiency levels for disaggregated models. Rank correlations between the overall efficiency results of different models vary from 0.99 to 0.61. General conclusions from all models are similar, but results for individual banks may change dramatically with different model specification. Conclusions from the model 1, which has been used in the analysis above, can thus be regarded as quite robust as far as bank groups or local banking sector as whole are concerned.

Table 6.1 Average results of different model specifications.[9]

Model	Technical efficiency	Scale efficiency	Overall efficiency
1	0.841	0.928	0.777
2	0.810	0.939	0.759
3	0.816	0.960	0.781
4	0.792	0.952	0.753
5	0.800	0.925	0.736
6	0.765	0.936	0.712
7	0.759	0.933	0.704
8	0.744	0.952	0.706

DEA is often criticised for not being able to cope with non-trivial random errors, which, if they exist, can have serious implications to the interpretations of the results. If random errors due to e.g. measurement errors occur in a best practise bank, the construction of the efficient frontier would be affected. On the contrary, errors that occur in an inefficient bank are not critical, since these errors affect only the efficiency ranking of the particular bank. Sensitivity of the relatively efficient frontier can be addressed by altering the data set. If some or all frontier observations are deleted from the data and the analysis is performed again sensitivity can be checked by examining the rank correlation between the old and new efficiency scores. If the data contains influential outliers among the frontier banks the rank correlation should be low. Although, the use of the pooled data will probably reduce the effects of random variation, frontier sensitivity is tested by deleting all frontier banks from the pooled data and re-running the analysis. Table 6.2 shows the rank correlation results for the original and new technical efficiency scores.

[9] Inputs in all models are labour, operating costs, and machinery.
Outputs in the models are:

Disaggregated models:	Aggregated models:
Model 1: SL, LL, DP, CA, NB, OE	Model 5: LO, DE, NB, OE
Model 2: SL, LL, DP, CA, OE	Model 6: LO, DE, OE
Model 3: SL, LL, DP, CA, NB,	Model 7: LO, DE, NB
Model 4: SL, LL, DP, CA	Model 8: LO, DE

SL = short term loans, **LL** = long term loans, **DP** = deposits by the public, **CA** = cheque accounts by the public, **NB** = number of branches, **OE** = other earnings, **LO** = loans aggregated (SL+LL), **DE** = deposits aggregated (DP+CA).

Table 6.2 **Frontier sensitivity results**

Number of banks deleted	Rank correlation	Number of frontier banks
None	1	CRS: 38 VRS: 76
CRS:38	0.93	62
VRS: 76	0.80	193

When all original frontier banks are deleted from the data the rank correlation in CRS-case is 0.93 and in VRS-case 0.80. The number of new frontier banks is in CRS-case 62 and in VRS-case 193. Thus, fairly close to the original VRS frontier there seems to be a bulk of banks that pop up to the frontier when data is altered. The average increase in technical efficiency was about 5 % in the case of new frontier and standard deviation of efficiency changes was quite small. This suggests that the relative frontiers are rather stable and thus the results can be considered quite robust to possible random errors in the frontier banks. It seems that possible random errors in the data are not very large and therefore DEA is a reliable method (see Banker et. al., 1993). Needless to say, this kind of sensitivity testing says nothing about possible errors that would result downward biased efficiency score for a bank. But since the main interest here is the best practice frontier, the possible negative errors are not examined.

6.2.6 Effects of mergers on efficiency

During 1986–1989 the number of Finnish local banks reduced from 624 to 509. 10 mergers took place in cooperative banks and 64 in savings banks. Average efficiency scores of the merged and original banks are presented in table 6.3.

Table 6.3 **Average efficiency scores for real and constructed observations from the pooled data**

	Technical efficiency (CRS)	Technical efficiency (VRS)	Scale efficiency	Sum of λ's
All observations	0.777	0.841	0.928	2.8
Original banks:	0.786	0.837	0.942	2.1
Cooperative banks	0.791	0.842	0.942	1.9
Savings banks	0.771	0.822	0.941	2.6
Merged banks:	0.671	0.898	0.750	11.3
Cooperative banks	0.768	0.912	0.844	4.3
Savings banks	0.637	0.893	0.717	13.9

Evaluation of success of the mergers with respect to technical efficiency depends heavily on the assumption about technology. Technical efficiency with constant returns to scale technology shows poor performance for the merged banks, especially for the merged savings banks. On the contrary, in the case of variable returns to scale technical efficiency is on average higher in merger banks. Better technical efficiency with VRS is, however, offset by large scale inefficiencies, as is to be expected from the previous analysis.

Scale inefficiencies of the merged banks even dominate their technical inefficiencies, which is opposite to the results of the non-merger banks. The average sum of λ's indicates that scale inefficiency, in particular in savings banks, is clearly due to decreasing returns to scale. Almost all of the merged banks belong to the largest size quartile, which explains the nature of their scale properties (see figure 6.12). The effect of scale inefficiency is exceptionally strong in savings banks. The non-merged banks are on average more efficient than merged ones. However, the mergers of cooperative banks seem to have been relatively more successful than those of the savings banks.

A common feature in the mergers of local banks has been that the acquiring bank has been quite large and the acquired bank quite small. These kinds of mergers have impact only on local market structure and clear distinction must be made between these 'small' mergers and mergers of whole banking groups or large commercial banks. Also, as Berger and Humphrey (1992) suggest, the mergers of 1990's may be more likely to bring about efficiency gains than the mergers of the 1980's because merger participants are now more motivated to try to reach efficiency gains. In the 1980's mergers were probably more motivated by size and market expansion, while now the interest is in cost reduction and removal of inefficient management.[10]

Even though recent literature has not on average found significant efficiency gains generated by mergers, it does not exclude the possibility of successful mergers. Almost all merger studies have found that at least some individual mergers have improved the performance of the merged banks profoundly. Unfortunately, the studies have not been able to find out what are the factors resulting a successful merger. Most of the research has concerned with input inefficiencies. However, as indicated by Berger et al. (1993) output inefficiencies may be as important or even more important than overuse of inputs. Therefore it is possible that mergers improve efficiency, but only on the output side. By merging two banks might be able to achieve a higher revenue output bundle through improved marketing, product innovation, repricing, risk management, or other revenue-enhancing effects (Berger, Hunter and Timme, 1993).

6.3 Productivity growth

For the study of productivity growth the bank structure of year-end 1990 is used. The banks that merged sometime during 1985–1990 are added together so that they appear as one bank for the whole period. Data sets for years in 1985–1989 thus include some observations that are an aggregate of banks that later merged. Table 6.3 shows the data of the average bank for this "merger corrected" data. The data for year 1990 is the same which was used above in the efficiency analysis. As can be noticed from the last column of the table 6.3 the percentage growth of the variables during 1985–1990 differ considerably. The most striking change has taken place in the input usage. The use of labour has grown only by 5 % while other operating expenses and machinery and equipment have more than doubled.

Table 6.3 **Data for average bank**

	1985	1986	1987	1988	1989	1990	%-change in 85–90
Number of personnel	40	41	42	43	44	42	5 %
Operating expenses	3941	4599	5555	6861	7667	8083	105 %
Machinery and equipment	683	788	933	1043	1352	1470	115 %
Total loans	136266	148927	169260	208130	231730	239407	76 %
Total deposits	141525	150990	163966	191691	198718	200192	41 %
Other earnigs	2763	3275	3824	6938	4949	5486	99 %

On the output side the total amount of loans has grown faster than the total amount of deposits. This indicates that the structure of funding the loans has changed. In particular, the growth of long term loans has been faster than the growth of short term loans. In the class of total deposits cheque accounts have grown more than other deposits by the public. The growth of the last output indicator, other earnings, has been well above of other output variables. It includes different fees and commissions for example from payment processing, security markets and currency exchange, provisions from bank guarantees, earnings from real estates and also earnings from sales of various properties. Rapid growth of commissions and fees is at least partly an outcome of deregulation. The sales of various properties is somewhat random in nature and may therefore cause some bias in productivity measurement. This is the case especially in savings banks in 1988 when more than half of the other earnings came from sales of properties. However, in 1990 the percentage of earnings from property sales relative to the aggregate of other earnings is on the same level as in 1985 and, therefore, it should not create problems when the whole period is studied.

6.3.1 The factor productivity ratios

Table 6.4 shows how the ratios between the chosen inputs and outputs have developed. These ratios cannot be used to study total productivity growth, but they indicate some trends in the input usage and development of the output structure. According to the ratios the labour productivity would have increased and productivity of capital decreased. This is not necessarily correct. As

[10] The alternative motives of bank mergers are discussed in detail by Hawawini and Swary (1990, 23–36).

mentioned, there has been a shift in technology towards growing use of machinery and equipment. One reason for this is that relative prices have changed in favour of capital in the later half of 1980's. Routines that previously required much labour time have been nearly completely automated. The nature of the labour input has changed and therefore simple ratios are inadequate for reliable study of productivity growth. Same is true for the productivity of capital. Use of machinery and equipment has extended to handle wider range of tasks than before and hence productivity measurement with simple input-output ratio is insufficient.

Table 6.4 **Factor productivity ratios for an average bank**

	1985	1986	1987	1988	1989	1990	%-change in 85–90
Loans / personnel	3418	3649	4053	4844	5239	5714	67 %
Deposits / personnel	3550	3700	3926	4462	4493	4778	35 %
Other earnigs / personnel	69	80	92	161	112	131	89 %
Loans / op. expenses	35	32	30	30	30	30	-14 %
Deposits / op. expenses	36	33	30	28	26	25	-31 %
Other earnings / op. expenses	0.70	0.71	0.69	1.01	0.65	0.68	-3 %
Loans / machinery	200	189	181	200	171	163	-18 %
Deposits / machinery	207	192	176	184	147	136	-34 %
Other earnings / machinery	4.05	4.15	4.10	6.65	3.66	3.73	-8 %

Of the three output indicators the grown importance of other earnings can be noticed against each input indicator. It has grown more than loans or deposits against labour and dropped less than loans or deposits against operating expenses or machinery and equipment. Operating expenses have grown rapidly mostly due to increased use of technology. Especially rents and leases, ADP-expenses, and expenses on real estates increased considerably from 1985 to 1990. In general, it seems that input usage has been affected by technological advancement and the output structure has changed because of the new environment caused by deregulation.

6.3.2 Total productivity growth

Malmquist indices were calculated for the period from 1985 to 1990. Development from year to year was examined with successive reference technologies. In addition, years 1985 and 1990 were compared directly. Input vector consisted of the same variables as above with efficiency analysis (labour, operating expenses, machinery and equipment). From the output vector number of branches was left out and only loans (short and long term loans separately), deposits (cheque accounts and deposits by the public separately) and other earnings were used. Number of branches was excluded since the

variable is fairly stable in time and also because the role of branches has changed after introduction of atm's and home banking facilities.

Technological progress shifts the constructed production frontier upwards and changes in input and output compositions alter the form of the frontier. As seen above, both input mix and output mix of the local banks changed dramatically during 1985–1990. This causes the frontiers of 1985 and 1990 to intersect and therefore Malmquist indices relative to 1985 technology give different results than indices relative to 1990 technology. Here the technology of 1985 is chosen as main reference, since it seems more reasonable to compare the present to the past than the other way around.

Table 6.5 **Total productivity indices of an average bank**

	All banks		Cooperative banks	Savings banks
Period (i→j)	$M_i(i,j)$	$M_j(i,j)$	$M_i(i,j)$	$M_i(i,j)$
1985–1986	1.013	0.957	1.009	1.023
1986–1987	1.018	0.983	1.019	1.015
1987–1988	1.056	1.000	1.050	1.069
1988–1989	1.043	1.027	1.058	1.005
1989–1990	1.046	1.020	1.039	1.064
Cumulated index 1985–1990	1.188	0.985	1.187	1.188
	1.288	0.906	1.261	1.352

Table 6.5 shows the total productivity indices for the average bank. From 1985 to 1990 productivity growth relative to 1985 technology was 29 % in an average local bank. For average cooperative bank total productivity growth was 26 % and for average savings bank 35 %. Higher Malmquist index scores for savings banks must be at least partly due to faster credit expansion in savings banks than in cooperative banks. However, this analysis does not include any indicator of e.g. the quality of the granted loans, and thus the indices must be interpreted with some care.

In periods 1985–1986, 1986–1987 and 1985–1990 the two indices, M_i and M_j, give dissimilar results relative to different technologies. This is due to intersection of the frontiers, which is caused by a change in input and output structures of the average bank. From 1987 to 1990 productivity growth seems to have been quite stable around 4–5 % per year. Index of savings banks varies more than index of cooperative banks.

6.3.3 Frontier productivity growth and catching up indices

Malmquist total productivity index can be decomposed into the effect of frontier's shift i.e. technological progress and into effect of the change in relative efficiency. Table 6.6 presents results for the frontier productivity growth and table 6.7 for the 'catching up' efficiency. When the whole period from 1985 to 1990 is examined total productivity growth in an average bank is

found to be totally due to technological progress. Relative efficiency of all local banks was on average the same in 1985 and in 1990. However, some differences can be observed between cooperative and savings banks. Technological progress seems to have been faster in cooperative banks but their relative efficiency has slightly dropped. In savings banks, on the other hand, frontier productivity growth is below the average of all local banks but the 'catching up' index shows about 8 % increase in relative efficiency. It must once again be noted that efficiency improvement in savings banks is probably to some extent caused by fast credit expansion, which later created huge credit losses.

Frontier productivity indices indicate that the period 1985–1986 seems to be a step back in terms of technology while progress can be observed from 1987 onwards. Average changes in relative efficiency in either direction, on the other hand, have been quite modest. The average catching up index of all local banks improved a little in the first two years, dropped slightly in 1988, and after that stayed stable. Effect of fast credit expansion on relative efficiency in savings banks gets some support when relative efficiency changes in cooperative and savings banks are compared. During the period 1985–1987 the relative efficiency of cooperative banks increased while that of savings banks decreased. On the contrary, during the expansion years 1988–1990, the relative efficiency decreased in cooperative banks and increased in savings banks.

Table 6.6 **Frontier productivity indices of an average bank**

	All banks		Cooperative banks	Savings banks
Period (i→j)	$FR_i(i,j)$	$FR_j(i,j)$	$FR_i(i,j)$	$FR_i(i,j)$
1985–1986	0.989	0.935	0.998	0.967
1986–1987	1.000	0.965	1.000	1.004
1987–1988	1.077	1.019	1.082	1.067
1988–1989	1.041	1.028	1.045	1.033
1989–1990	1.053	1.026	1.065	1.025
Cumulated index 1985–1990	1.169	0.969	1.200	1.097
	1.287	0.900	1.300	1.256

Table 6.7 **Relative efficiency changes in an average bank**

Period	All banks	Cooperative banks	Savings banks
1985–1986	1.027	1.013	1.062
1986–1987	1.020	1.022	1.013
1987–1988	0.981	0.971	1.004
1988–1989	1.000	1.011	0.975
1989–1990	0.995	0.977	1.039
Cumulated index 1985–1990	1.023	0.994	1.094
	1.003	0.972	1.080

The figures 6.13–6.15 show the productivity growth percentage of each individual bank plotted against its total sum of loans during 1985–1990. 87 % of all local banks experienced at least some productivity growth (figure 6.13). Large banks seem to have increased productivity more than small banks. Almost all of the banks that show productivity decline are smaller in size than the average bank. The same trend can be noticed in figure 6.14 which presents frontier productivity growth percentages of individual banks. Technology regressed in only 6 % of the banks and most of them were small ones. The average relative efficiency remained on same level through 1985–1990 but some individual banks experienced large fluctuations (figure 6.15). About half of the banks increased and the other half decreased their relative efficiency. There, however, is no substantial difference between small and large local banks.

Figure 6.13 **Total productivity growth in individual banks 1985–1990. 1985 as reference technology.**

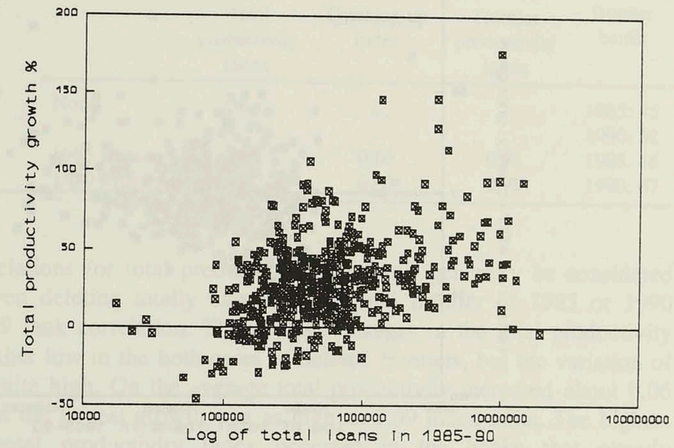
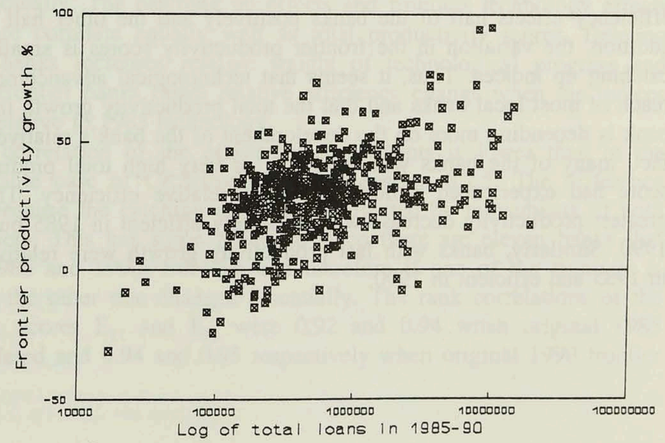
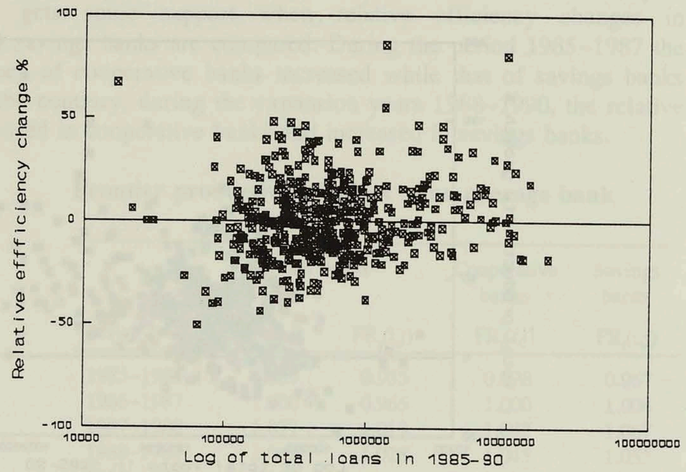


Figure 6.14 **Frontier productivity growth in individual banks 1985–1990. 1985 as reference technology**



Some reasons for higher productivity growth rates in large banks are intuitive. It seems clear that large banks would adapt new technology faster and more easily than small banks and hence productivity growth would be greater. Also, in large banks the capacity of machinery and equipment can be utilised more efficiently. For example in ADP routines there probably exists some degree of economies of scale. In the case of very small banks low productivity growth might be due to scale inefficiency which is caused by increasing returns to scale. Small banks may also have relatively more excess capital than large banks. Moreover, since the trend in input usage is towards automation and exploitation of machinery, it would feel that very small banks might have problems in keeping up with technological progress that speeds productivity growth.

Figure 6.15 **Changes in relative efficiency in individual banks 1985–1990**



According to figure 6.13 almost all of the banks have been able to utilize technological progress at least to some extent. Instead, the change in relative efficiency effects half of the banks positively and the other half negatively. In addition, the variation in the frontier productivity scores is smaller than in the catching up indices. Thus, it seems that technological advancement is within a reach of most local banks and that the total productivity growth in an individual bank is depending more on the development of the bank's relative efficiency. In fact, many of the banks with very low or very high total productivity growth score had experienced radical change in relative efficiency. The banks with greatest productivity decrease were relatively efficient in 1985 but inefficient in 1990. Similarly, banks with fast productivity growth were relatively inefficient in 1985 and efficient in 1990.

6.3.4 Sensitivity of productivity scores

The above Malmquist index results can be thought a priori to be more sensitive to measurement errors than the results of the plain efficiency analysis of the section 6.2. The reason is mostly that the data set compared have to be single cross-sections and thus the advantages of using a pooled data are not present. The robustness of the results for the period 1985–1990 was examined similarly as above i.e. by deleting the original frontier banks and then checking the rank correlation between the old and the new Malmquist indices. Table 6.8 shows the correlation outcome.

Table 6.8 **Sensitivity analysis of the Malmquist indices relative to 1985 technology**

Deleted banks	Rank correlations			Number of frontier banks
	Total productivity index	Catching up index	Frontier productivity index	
None	1	1	1	1985: 45 1990: 32
1985 frontier	0.91	0.86	0.82	1985: 46
1990 frontier	0.92	0.88	0.71	1990: 67

The rank correlations for total productivity scores can generally be considered quite high. Even deleting totally either the original frontier of 1985 or 1990 results over 0.9 rank correlation. The average changes in the total productivity scores were rather low in the both cases of deleted frontiers, but the variation of changes was quite high. On the average total productivity increased about 0.06 index units, but the largest growth was as high as 1.09 index units. The highest increases in total productivity index occurred in the banks that already originally had relatively high index value.

Because the two components of total Malmquist index respond differently to deletion of one of the frontiers, is obvious that some individual index scores may change radically. The catching up effects and frontiers technology effects do not therefore correlate equally well as total productivity scores. Deleting 1985 frontier banks increases relative weight of technological progress and deleting 1990 frontier banks raises relative efficiency change when the indices are measured relative to 1985 technology.

The rank correlation of the efficiency components is lower than in the almost comparable case of CRS-frontier in the section 6.2.5. The difference raises mostly because the catching up index is affected by movements of both reference frontiers. This has some influence since there are eleven banks that lay on both, 1985 and 1990, frontiers and therefore if one of the frontiers is totally deleted, the other one changes essentially. The rank correlations of the plain efficiency scores E_{11} and E_{22} were 0.92 and 0.94 when original 1985 frontier was deleted and 0.94 and 0.95 respectively when original 1990 frontier

was deleted. The individual CRS-frontiers can thus be considered quite stable relative to random errors.

The calculated Malmquist total productivity scores for the whole period 1985–1990 were found to slightly correlate with the share of loan losses in 1991 relative outstanding loans in that period.[11] This means that the estimated productivity growth rates might be somewhat biased because of expansion of poorly evaluated loans. The problem concerns especially large local banks that grew very fast during the last years of 1980's. The Malmquist indices were recalculated for the period 1985–1990 so that experienced loan losses of 1991 were subtracted from the loans of 1990. Average levels of new results were similar to previous ones. The largest changes occurred at the largest banks so that the trends in the figures 6.13 and 6.14 moderately flattened. Therefore, the actual productivity growth differences between the smallest and the largest classes of local banks are not necessarily as large as the section 6.3.3 suggests and the results must be viewed critically.

6.3.5 Effects of mergers on productivity

Data used in the productivity analysis consisted of 471 real observations and 38 constructed merger observations. Table 6.9 shows the average productivity indices of the real and constructed banks.

Table 6.9 **Malmquist indices of the real and constructed observations in period 1985–1990 relative to 1985 technology**

	Total productivity	Catching up	Frontier productivity
All observations	1.288	1.003	1.287
Original banks:	1.282	0.999	1.284
Cooperative banks	1.257	0.972	1.295
Savings banks	1.352	1.079	1.256
Merged banks:	1.365	1.051	1.310
Cooperative banks	1.401	0.964	1.463
Savings banks	1.353	1.082	1.255

The results generally indicate that productivity growth would have been faster in merger banks than in non-merger banks. Especially striking the difference is in cooperative banks, where the average productivity growth of the merger banks was 40 % from 1985 to 1990. This was all due to technological progress component. Similar trend cannot be observed in savings banks, where the index values of real and constructed banks are analogous. These rough averages thus suggest that mergers in cooperative banks would have been more successful than the merger in savings banks.

[11] Correlation was 0.156, which is significant at the 5 % level but not at the 0.1 % level.

7 Conclusions

This study is concerned with measuring and explaining producer performance in Finnish local banking. Primary interest is in the production of retail banking services. Performance is viewed as a function of the state of technology and economic efficiency. The former defines a frontier relation between inputs and outputs and the latter incorporates waste and misallocation relative to this frontier. Improvements in bank performance can occur through innovation in technology and through increase in efficiency.

The main result of the efficiency analysis is that technical inefficiencies dominate scale inefficiencies in Finnish local banking. Average level of technical inefficiency during 1985–1990 was little less than 16 %. Average level of scale inefficiency on its half was only about 7 %. Distributions of technical inefficiencies were similar for cooperative and savings banks, but savings banks were found to be slightly more scale inefficient. However, fast credit expansion in the last years of 1980's that later created huge loan losses may bias efficiency evaluation especially in case of savings banks. Therefore, savings banks' efficiency scores must be viewed critically.

Since it is possible that best practice banks are continuously about 20 % more cost efficient than an average bank, it appears that banks are either producing services with very different quality or there is not much competitive pressure to control costs. The latter seems more plausible considering the situation of the local banking market in the late 1980's. The large and persistent cost efficiency differences between banks of similar size and product mix suggest that greater competition within the banking industry would be beneficial.

Data indicates that the way we choose to measure bank production influences especially the efficiency rankings between individual banks. The level of measured inefficiency depends somewhat of the level of aggregation and number of used variables. However, relation between technical and scale inefficiencies are similar in all variable specifications and average results of the efficiency analysis are robust to model specification errors and to effect of random errors.

There are potential productive efficiency gains to be achieved in Finnish local banking by improving the level of technical efficiency. Over 60 % of loans were produced in banks that were technically inefficient. Technical inefficiency does not appear to be closely related to bank size. Potential benefits of scale efficiency improvements in the local banking sector are much smaller. Scale inefficiency was found mostly to be due to decreasing returns to scale. Furthermore, large banks were more likely to be scale inefficient than small banks and thus optimal size for a local bank appears to be quite small. However, large local banks are probably more involved in e.g. money market activities, which requires input usage that is included in the input vectors of this study but the output of these kind of activities is not necessarily fully represented.

Overall efficiency scores were on average lower in merged banks than in non-merged banks. Thus, the most important motivational factor leading to mergers in the 1980's was probably not cost reduction. Inefficiency of the merged banks was mostly due to scale inefficiency since at least one party of a

merger was usually a large bank. On average merged cooperative banks were found more efficient than merged savings banks.

The main result of the productivity analysis was that average productivity growth was found to be totally due to technological progress rather than improvements in relative efficiency. Almost all banks were able to achieve some technological progress. Even though relative efficiency changes were on average close to zero, the banks that had largest changes in total productivity were those that had very large changes in relative efficiency. This suggests that the most important factor influencing individual bank's total productivity growth is efficiency. Measured average total productivity growth rates in the late 1980's varied from 1 % to 5 % annually. Productivity growth was found to be on average faster in merged banks than in non-merged banks. However, production of bad loans may bias Malmquist indices to show too high productivity growth especially in large merger banks.

References

Aly, H.Y., Grabowski, R., Pasurka, C., Rangan, N. (1990) Technical, scale, and allocative efficiencies in U.S. banking: An empirical investigation, Review of Economics and Statistics 72, 211–218.

Banker, R.D., Charnes, A., Cooper, W.W. (1984) Some models for estimating technical and scale inefficiencies in data envelopment analysis, Management Science 30, 1078–1092.

Banker, R.D., Gadh, V.M., Gorr, W.L. (1993) A Monte Carlo comparison of production frontier estimation methods: Corrected ordinary least squares and data envelopment analysis, European Journal of Operational Research 67, 332–343.

Bauer, P.W., Berger, A.N., Humphrey, D.B. (1993) Efficiency and productivity growth in U.S banking, in: H.O. Fried, C.A.K. Lovell and S.S. Schmidt eds., The measurement of productive efficiency: Techniques and applications, Oxford University Press, 386–413.

Berg, S.A. (1992) Mergers, efficiency and productivity growth in banking: the Norwegian experience 1984–1990, Norges Bank, Arbeids Notat 1992/3, June.

Berg, S.A., Claussen, A.C., Førsund, F.R. (1993) Banking efficiency in the Nordic countries: A multi-output analysis, Norges Bank, Arbeids Notat 1993/3, March.

Berg, S.A., Førsund, F.R., Jansen, E.S. (1989) Bank output measurement and construction of best practice frontiers, Norges Bank, Arbeids Notat 1989/6, July.

Berg, S.A., Førsund, F.R., Jansen, E.S. (1992) Malmquist Indices of Productivity Growth during the Deregulation of Norwegian Banking 1980–1989. Scandinavian Journal of Economics 94, Supplement, 211–228.

Berger, A.N. (1991) The profit-concentration relationship in banking. Finance and Economics Discussion Series paper no. 176, November, Federal Reserve Board, Washington, D.C.

Berger, A.N. (1992) "Distribution free" estimates of efficiency in the U.S. banking industry and tests of the standard distributional assumptions. Finance and Economics Discussion Series paper no. 188, March, Federal Reserve Board, Washington, D.C.

Berger, A.N., Hancock, D., Humphrey, D.B. (1993) Bank efficiency derived from a profit function, Journal of Banking and Finance 17, 317–347.

Berger, A.N., Humphrey, D.B. (1990) Measurement and efficiency issues in commercial banking, Finance and Economics Discussion Series paper no. 151, December, Federal Reserve Board, Washington, D.C.

Berger, A.N., Humphrey, D.B. (1991) The dominance of inefficiencies over scale and product mix economies in banking, Journal of Monetary Economics 28, 117–148.

Berger, A.N., Humphrey, D.B. (1992) Megamergers in banking and the use of cost efficiency as an antitrust defence. Finance and Economics Discussion Series, paper no. 203, June, Federal Reserve Board, Washington, D.C.

Berger, A.N., Hunter, W.C., Timme, S.G. (1993) The efficiency of financial institutions: A review and preview of research past, present, and future, Journal of Banking and Finance 17, 221–249.

Button, K.J., Weyman-Jones, T.G. (1992) Ownership Structure, InstitutionalOrganisation and Measured X-Efficiency, American Economic Review 82, No. 2, 439–445.

Caves, D.W., Christensen, L.R., Diewert, W.E. (1982) The economic theory of index numbers and the measurement of input, output and productivity. Econometrica 50, 1393–1414.

Charnes, A., Cooper, W. (1962) Programming with linear fractional functionals, Naval Research Logistics Quarterly 9, 181–185.

Charnes, A., Cooper, W., Golany, B., Seiford, L., Stutz, J. (1985) Foundations of Data envelopment analysis for Pareto-Koopmans efficient empirical production functions, Journal of Econometrics 30, 91–107.

Charnes, A., Cooper, W., Rhodes, E. (1978) Measuring the efficiency of decision making units, European Journal of Operational Research 2, 429–444.

Charnes, A., Cooper, W., Seiford, L., Stutz, J. (1983) Invariant multiplicative efficiency and piecewise Cobb-Douglas envelopments, Operations Research Letters 2, 101–103.

Colwell, R.J., Davis E.P. (1992) Output and productivity in banking, Scandinavian Journal of Economics 94, Supplement, 111–129.

Debreu, G. (1951) The coefficient of resource utilization, Econometrica 19, 273–292.

Elyasiani, E., Mehdian, S.M. (1990) A nonparametric approach to measurement of efficiency and technological change: The case of large U.S. commercial banks, Journal of Financial Services Research 4, 157–168.

Elyasiani, E., Mehdian, S.M. (1992) Productive efficiency performance of minority and nonminority-owned banks: A nonparametric approach, Journal of Banking and Finance 16, 933–948.

Färe, R., Grosskopf, S., Lindgren, B., Roos, P. (1989) Productivity developments on Swedish hospitals, A Malmquist output index approach. Mimeo.

Färe, R., Grosskopf, S., Lovell, C.A.K. (1985) The Measurement of Efficiency of Production, Kluwer-Nijhoff Publishing.

Färe, R., Lovell, C.A.K. (1978) Measuring the technical efficiency of production, Journal of Economic Theory 19, 150–162.

Farrell, M.J. (1957) The Measurement of productive efficiency, Journal of the Royal Statistical Society A 120, 253–281.

Ferrier, G.D., Lovell, C.A.K. (1990) Measuring cost efficiency in Banking: Econometric and linear programming evidence, Journal of Econometrics 46, 229–245.

Forestieri, G. (1993) Economies of scale and scope in the financial services industry: A review of the literature, in: Financial conglomerates, OECD, Paris, 63–124.

Førsund, F.R. (1990) The Malmquist productivity index, Memorandum No. 28, Department of Economics, University of Oslo, December.

Førsund, F.R. (1992) The DEA programme for calculating efficiency, in: Leppänen, S., Loikkanen, H. (eds.) Proceedings of the workshop on the evaluation of public sector performance, VATT-publications 7, Government Institute for Economic Research, Helsinki, 29–39.

Førsund, F.R., Hjalmarsson, L. (1974) On the measurement of productive efficiency, Swedish Journal of Economics 76, 141–154.

Førsund, F.R., Hjalmarsson, L. (1979) Generalized Farrell measures of efficiency: An application to milk processing in Swedish dairy plants, Economic Journal 89, 294–315.

Førsund, F.R., Lovell, C.A.K., Schmidt, P. (1980) A survey of frontier production functions and of their relationship to efficiency measurement, Journal of Econometrics 13, 5–25.

Hawanini, G., Swary, I. (1990) Mergers and Acquasitions in the U.S. Banking Industry, Evidence from the Capital Markets. North-Holland.

Humphrey, D.B. (1990) Why do Estimates of Bank Scale Economies Differ? Federal Reserve Bank of Richmond, Economic Review 76 (September/October), 38–50.

Intriligator, M.D. (1971) Mathematical optimization and economic theory. Prentice-Hall Inc., Englewood Cliffs, N.J.

Kim, M., Weiss, J. (1989) Total Factor Productivity Growth in Banking: The Israeli Banking Sector 1979–1982. Journal of Productivity Analysis 1, 139–153.

Kolari, J., Zardkoohi, A. (1987) Bank costs, structure, and performance. Lexington Books, D.C. Heat and Company, Massachusetts.

Koopmans, T.C. (1951) An analysis of production as an efficient combination of activities, in: T.C. Koopmans (ed.): Activity analysis of production and allocation, Cowles Commission for Research in Economics, Monograph no. 13, New York, John Wiley and Sons Inc.

Liebenstein, H. (1966) Allocative efficiency vs. 'X-efficiency', American Economic Review 56, 392–415.

Liebenstein, H., Maital, S. (1992) Empirical Estimation and Partitioning of X-Inefficiency: A Data Envelopment Approach. American Economic Review 82, May, 428–433.

Lovell, C.A.K. (1993) Production frontiers and productive efficiency, in: H.O. Fried, C.A.K. Lovell and S.S. Schmidt eds., The measurement of productive efficiency: Techniques and applications, Oxford University Press, 3–67.

Malmquist, S. (1953) Index numbers and indifference surfaces. Trabajos de Estadistica 4, 209–242.

Mester, L. (1993) Discussants' comments on Berg et al. and McAllister and McManus, Journal of Banking and Finance 17, 407–409.

Nishimizu, M., Page, J.M. (1982) Total factor productivity growth, technological progress and technical efficiency change: Dimensions of productivity change in Yugoslavia, 1965–78, Economic Journal 92, 920–936.

Parkan, C. (1987) Measuring the efficiency of service operations: An application to bank branches. Engineering Costs and Production Economics 12, 237–242.

Rangan, N., Grabowski, R., Aly, H.Y., Pasurka, C. (1988) The technical efficiency of US banks, Economic Letters 28, 169–175.

Revell, J.R. (1980) Costs and Margins in Banking: An International Survey. OECD, Paris.

Schmidt, P., Sickles, R.C. (1984) Production frontiers and panel data, Journal of Business & Economic Statistics 2, 367–374.

Seiford, L.M., Thrall, R.M. (1990) Recent developments in DEA: The mathematical programming approach to frontier analysis, Journal of Econometrics 46, 7–38.

Sherman, H.D., Gold, F. (1985) Bank branch operating efficiency: Evaluation with data envelopment analysis, Journal of Banking and Finance 9, 297–315.

Appendix 1

A numerical example of linear programming problem

The first step in solving a linear programming problem is to add slack variables in order to eliminate inequalities. The BCC-model can then be written in the following form:

$$\text{Min}_{\theta,\lambda,s,e}\ z_0 = \theta - \varepsilon(s + e)$$

$$\begin{aligned} \text{s.t.}\quad & Y\lambda - s = Y_0 \\ & \theta X_0 - X\lambda - e = 0 \\ & I_v\lambda = 1 \\ & \lambda,s,e \geq 0 \end{aligned}$$

where θ is efficiency score, ε is non-Archimidean, s and e are the slack variables, λ is the weight vector, Y is output vector, X is input vector and I_v is a unit vector.

Example data:	DMU	X	Y
	1	2	2
	2	3	5
	3	6	7
	4	5	3
	5	4	1

The linear problem then e.g. for DMU 4 is

$$\text{BCC}(\text{DMU}_4):\quad \text{Min}_{\theta,\lambda,s,e}\ \theta - \varepsilon(s + e)$$

$$\begin{aligned} \text{s.t.}\quad & 2\lambda_1 + 5\lambda_2 + 7\lambda_3 + 3\lambda_4 + \lambda_5 - s = 3 \\ & \theta 5 - 2\lambda_1 - 3\lambda_2 - 6\lambda_3 - 5\lambda_4 - 4\lambda_5 - e = 0 \\ & \lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5 = 1 \end{aligned}$$

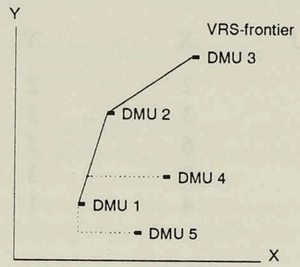
This problem is solved in two stages, first θ and then the slack variables. There are three independent constraints with five weight variables, two slack variables and the efficiency score. A basic feasible solution is a vector of λ's such that any three of them are set equal to 0 and the remaining to are non-negative and satisfy the above constraints. The problem is then solved by iterating through different combinations of λ's and the solution that leads to minimum θ is optimal. The slack variables are then solved with the optimal λ's and the θ. The problem of the CCR-model is similar as above except that the last constraint is omitted. The solution is iterated analogously.

The results for example data are:

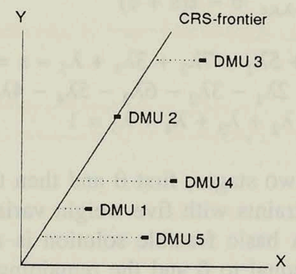
	BCC				CCR			
DMU	θ^*	s	e	λ	θ^*	s	e	λ
1	1	0	0	$\lambda 1 = 1$	3/5	0	0	$\lambda_2 = 0.4$
2	1	0	0	$\lambda_2 = 1$	1	0	0	$\lambda_2 = 1$
3	1	0	0	$\lambda_3 = 1$	7/10	0	0	$\lambda_2 = 1.4$
4	7/15	0	0	$\lambda_1 = 2/3$ $\lambda_2 = 1/3$	9/25	0	0	$\lambda_2 = 0.6$
5	1/2	1	0	$\lambda_1 = 1$	3/20	0	0	$\lambda_2 = 0.2$

Graphically the constructed frontiers look like:

BCC-model:



CCR-model:



The nature of DMU's scale properties can be revealed by studying the sum of λ's in CCR-model (see section 4.2). Unit 2 is the only scale efficient observation, since the sum of λ's is equal to one ($\lambda_2 = 1$). Units 1, 4, 5 experience increasing returns to scale, since their sum of λ's is below one. Unit 3 experiences decreasing returns to scale ($\lambda_2 = 1.4$).

Appendix 2

Data for an average savings and cooperative bank

Data for the average local bank 1985–1990 (in 1985 prices):

	1985	1986	1987	1988	1989	1990
Number of personel	35	36	38	40	44	42
Operating expenses	3415	4050	5012	6340	7548	8083
Machinery and equipment	694	825	978	1100	1474	1470
Short term loans	10891	11090	11529	13726	17792	18488
Long term loans	105763	118935	139415	179600	209729	220919
Cheque accounts	9738	9921	11229	14287	17893	21333
Deposits by the public	112720	123269	136157	163552	177607	178859
Other earnings	2467	2964	3582	6432	4972	5486
Number of deposits	13062	13455	13982	15125	15721	14736
Number of loans	3285	3435	4040	4802	5072	4882
Number of branches	4.1	4.2	4.2	4.4	4.7	4.5

Data for the average cooperative bank 1985–1990 (in 1985 prices):

	1985	1986	1987	1988	1989	1990
Number of personel	26	27	28	28	29	29
Operating expenses	2597	3013	3666	4497	4994	5357
Machinery and equipment	344	405	491	526	705	812
Short term loans	8892	8814	8772	9660	12344	12950
Long term loans	87826	96646	110355	134725	142462	146326
Cheque accounts	7281	7390	8261	10226	13611	16861
Deposits by the public	86913	93811	101919	117784	121171	123555
Other earnings	1717	2055	2339	3057	3194	3208
Number of deposits	9951	10121	10282	10742	10345	9661
Number of loans	2417	2487	2956	3215	3285	3359
Number of branches	3.3	3.3	3.3	3.3	3.3	3.2

Data for the average savings bank 1985–1990 (in 1985 prices):

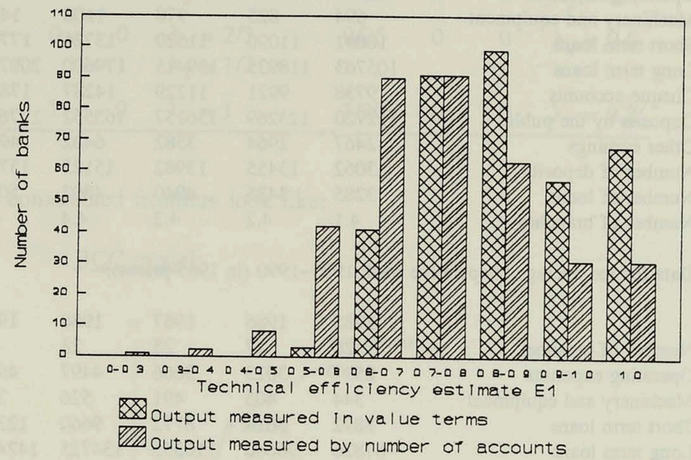
	1985	1986	1987	1988	1989	1990
Number of personel	47	50	54	61	73	74
Operating expenses	4608	5643	7172	9545	12726	14606
Machinery and equipment	1203	1471	1758	2097	3033	3044
Short term loans	13803	14584	15952	20797	28841	31743
Long term loans	131892	153155	186037	257653	346151	399447
Cheque accounts	13318	13808	15991	21352	26577	32038
Deposits by the public	150314	168496	191086	243159	292062	311221
Other earnings	3560	4360	5577	12302	8581	10938
Number of deposits	17592	18574	19919	22749	26623	26883
Number of loans	4551	4890	5780	7561	8698	8527
Number of branches	5.2	5.5	5.8	6.2	7.3	7.6

	1985	1986	1987	1988	1989	1990
GDP-deflator:	1.0000	1.0454	1.1010	1.1774	1.2578	1.3227
'Machinery' deflator (prices of machinery and equipment)	1.0000	1.0290	1.0603	1.0922	1.1400	1.1839
'Other costs' deflator (prices of materials)	1.0000	1.0290	1.0817	1.1373	1.2480	1.3239

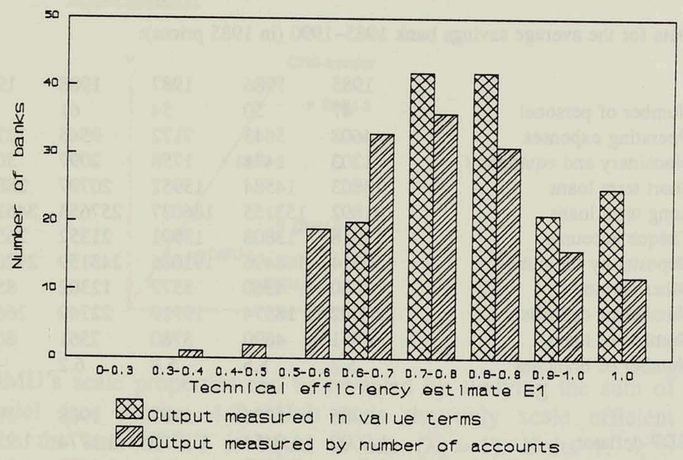
Appendix 3

Frequency distributions of technical and scale efficiency estimates for savings and cooperative banks

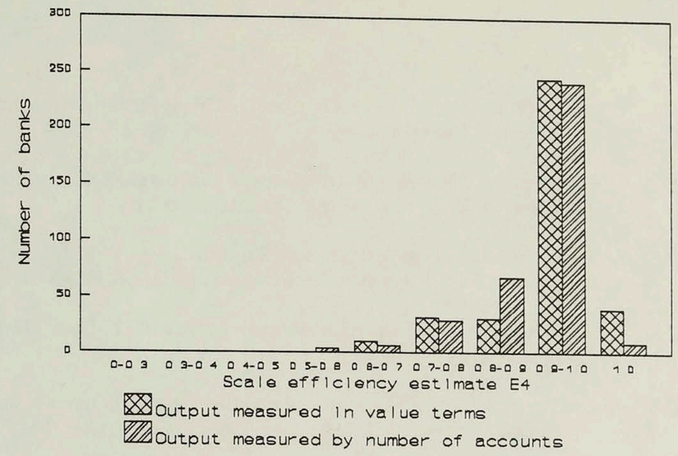
Frequency distribution of technical efficiency scores. Pooled data 1985–1990, cooperative banks



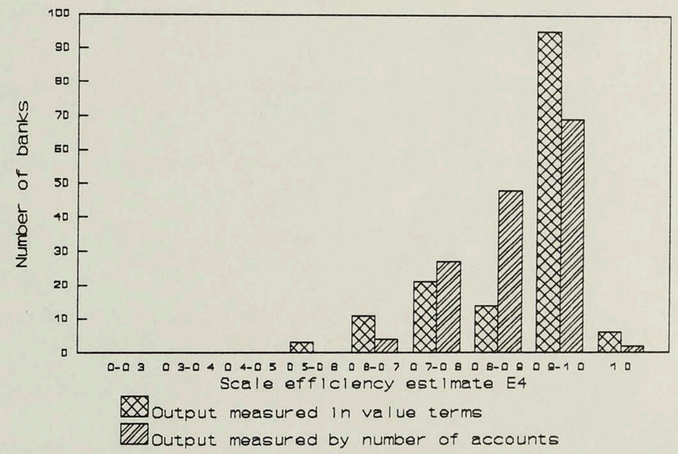
Frequency distribution of technical efficiency scores. Pooled data 1985–1990, savings banks



Frequency distribution of scale efficiency scores. Pooled data 1985–1990, cooperative banks



Frequency distribution of scale efficiency scores. Pooled data 1985–1990, savings banks



SUOMEN PANKIN KESKUSTELUALOITTEITA

ISSN 0785-3572

1/93 Shumin Huang **Determinants of Country Creditworthiness: An Empirical Investigation, 1980—1989**. 1993. 57 s. ISBN 951-686-363-9. (TU)

2/93 Rami Hakola **Pääoma- ja yritysverouudistuksen vaikutukset teollisuuden rahoitusrakenteeseen**. 1993. 45 s. ISBN 951-686-364-7. (KT)

3/93 Pentti Forsman — Pertti Haaparanta — Tarja Heinonen **Waste Paper Recycling and the Structure of Forest Industry**. 1993. 20 s. ISBN 951-686-365-5. (KT)

4/93 Risto Murto **Pankkiluottojen hinnoittelu vuosina 1987—1992: Mikä meni vikaan?**. 1993. 33 s. ISBN 951-686-366-3. (RM)

5/93 Johanna Pensala — Heikki Solttila **Pankkien järjestämättömät saamiset ja luottotappiot vuonna 1992**. 1993. 21 s. ISBN 951-686-367-1. (RM)

6/93 Harri Hasko **Valuuttakauppojen netotus ja riskien hallinta**. 1993 41 s. ISBN 951-686-368-X. (TU)

7/93 Jon Hirvilahti **Ensimmäisestä maailmansodasta toiseen kultakantaan**. Katsaus kelluvien valuuttakurssien ajanjaksoon vuosina 1914—1925. 1993. 120 s. ISBN 951-686-369-8. (TU)

8/93 Peter Nyberg — Vesa Vihriälä **The Finnish Banking Crisis and Its Handling**. 1993. 43 s. ISBN 951-686-370-1. (RM)

9/93 Anne Brunila — Kari Takala **Private Indebtedness and the Banking Crisis in Finland**. 1993. 39 s. ISBN 951-686-371-X. (KT)

10/93 Johanna Pensala — Heikki Solttila **Banks' Nonperforming Assets and Write-Offs in 1992**. 1993. 20 s. ISBN 951-686-372-8. (RM)

11/93 Sinimaaria Ranki **The ECU as the Future Currency of Financial Transactions**. 1993. 35 s. ISBN 951-686-373-6. (KP)

12/93 Matti Suominen **Fixed Rate Loan Contracts, Maturity Transformation and Competition in the Deposit Market**. 1993. 19 s. ISBN 951-686-374-4. (TU)

13/93 Esko Sydänmäki **EY:n instituutiot**. 1993. 35 s. ISBN 951-686-377-9. (KP)

14/93 Harri Kuussaari **Productive Efficiency in Finnish Local Banking During 1985—1990**. 1993. 67 s. ISBN 951-686-380-9. (TU)