Efficiency of financial institutions
Meesters, A.J.

IMPORTANT NOTE: You are advised to consult the publisher's version (publisher's PDF) if you wish to cite from it. Please check the document version below.

Document Version
Publisher's PDF, also known as Version of record

Publication date:
2009

Link to publication in University of Groningen/UMCG research database

Citation for published version (APA):

Copyright
Other than for strictly personal use, it is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license (like Creative Commons).

Take-down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Downloaded from the University of Groningen/UMCG research database (Pure): http://www.rug.nl/research/portal. For technical reasons the number of authors shown on this cover page is limited to 10 maximum.
Chapter 2

Efficiency

2.1 Introduction
This chapter explains the concept of efficiency and gives an overview of how efficiency can be estimated. Moreover it describes the choices that have to be made to estimate efficiency for the banking sector. Since the main methodology of this thesis is related to efficiency, it is important to make clear what efficiency is and what types of efficiency measures are developed. This chapter begins with a short introduction in which the history of the concept of efficiency is described. In the second section several types of efficiency are discussed. The third section of this chapter shows how efficiency can be estimated as well as the underlying assumptions. The fourth section explains how the theory can be used to analyze bank efficiency. Section five concludes.

Efficiency was introduced in the 1950s by Koopmans (1951).1 In his rather technical monograph Koopmans gives the definition of an efficient point: ‘A possible point [...] in the commodity space is called efficient whenever an increase in one of its coordinates (the net output of one good) can be achieved only at the cost of a decrease in some other coordinate (the net output of another good).’ In other words, a point is efficient if the output is maximized given the inputs. Debreu (1951) uses this definition to develop a measure of efficiency, or, in his own words: ‘A numerical evaluation of the “dead loss” associated with a non optimal situation (in the Pareto sense) of an economic system’. The general idea of this measure is to determine the distance between the produced output and the output that could have been produced given the inputs. Shephard (1953) uses the same concept of distance functions, yet, he states it as a problem that a producer uses too many inputs to

1 As far as I know this is the first publication that gives a definition of efficiency. ISI Web of Knowledge gives no older articles that deal with efficiency.
produce a certain amount of outputs. This means that Shephard has an input oriented view while Koopmans is looking at the output.

The work of Farrell (1957) uses the work of Koopmans (1951) and Debreu (1951). In this paper Farrell (1957) shows how distance functions can be used in a practical way. To illustrate this practical way he uses an empirical example of the efficiency in the agricultural sector. Strangely enough, the work of Shephard (1953) is neglected in the early literature on efficiency. The next section discusses the several types of efficiency and how the distance between produced output and optimal output can be measured.

2.2 Types of Efficiency

2.2.1 Technical efficiency

According to the definition of Koopmans (1951) a producer is efficient if she maximizes output given the input she uses. For transforming inputs to outputs one needs a certain kind of technology. Since this type of efficiency deals solely with technology this type of efficiency is called technical efficiency (TE). The following example will make clear how TE is measured. All the steps are graphically represented in Figure 2.1. Suppose that one input ($X$) is needed to produce two outputs ($Y_1$ and $Y_2$) by a certain technology. The simplest way to describe technology is by the use of a production function (Varian, 1992). A graph that represents this production possibility function is given by $S'S'$ in Figure 2.1. The x- and y-axis denote outputs $Y_1$ and $Y_2$ respectively. The curve $S'S'$ denotes the possibilities of output given an amount of $X$. In an ideal situation every producer who has that specific amount of input $X$ will produce somewhere on the curve $S'S'$. In a less ideal situation, however, it is possible that a producer produces less than the outputs represented by $S'S'$ for instance she produces $Y_1^*$ and $Y_2^*$ (represented by point $P$ in Figure 2.1). To determine TE a fully efficient point is necessary. Such a point is located on the curve $S'S$. Although it is possible to calculate the distance from point $P$ to each point on curve $S'S$, it makes more sense to choose a point that has the same characteristics as point $P$. This point is obtained by taking a point with the same ratio of $Y_1$ to $Y_2$ as $Y_1^*$ to $Y_2^*$. This point is represented in the Figure 2.1 by point $Q$. The distance between point $P$ and point $Q$ is now a measure for efficiency.
This measure has one drawback though. It is an absolute measure and it does not take into account what the amount of output is that could have been produced. E.g. a distance between point $P$ and $Q$ of one has a different interpretation if e.g. $Y_1$ and $Y_2$ are about ten or if they are about one million. In the former case the producer could produce significantly more, while in latter case the increase is marginal. To overcome this problem efficiency is a relative measure and can be determined by the ratio of distance $OP$ to $OQ$. This measure gives the value one if $P$ is equal to $Q$.

This is the case if the amount of outputs that are produced lie on the $S'S$ curve and thus is fully efficient. The measure gets a value of zero if $P$ is equal to $O$. This means that although inputs are used, no outputs are produced at all.

In the description of TE, as given above, output is maximized given the inputs. This type of TE is called the output oriented measure of TE. On the other hand, there also exists an input oriented measure of TE. This measure assumes that output is given and the producer minimizes her inputs. This measure is represented in Figure 2.2. Suppose that two inputs ($X_1$ and $X_2$) are needed to produce one output ($Y$) by a certain production process. A graph of the isoquant of this production process is made in Figure 2.2. The x-axis denotes input $X_1$ and the y-axis denotes input $X_2$. The curve $S'S$ represents the amounts of $X_1$ and $X_2$ to produce an identical amount of $Y$. This means that in an ideal situation every producer that wants to produce a certain amount of output $Y$ needs the amount of inputs represented by the frontier. In a less ideal situation it is possible that a producer needs $X_1^*$ and $X_2^*$ for
Chapter 2

the production of amount $Y$. This is represented by point $P$. To determine efficiency a full efficient point is necessary. Such a point is located on the curve $S'S$. Just as in the output oriented case the proportion of $X_1$ and $X_2$ should be equal to the proportion of $X_1^*$ and $X_2^*$. This point is represented by point $Q$. With the use of point $Q$ efficiency can be determined by the ratio of distance $OQ$ to $OP$. This measure gives the value one if $P$ is equal to $Q$. This is the case if the amount of inputs that are needed for the production lay on the $S'S$ curve and thus is fully efficient. The measure gets a value of zero if $P$ is equal to infinity.

In the examples given above, a measure for TE is shown for the output as well as the input oriented case. For the output oriented case one input and two outputs are assumed yet it also holds in the case of multiple inputs and multiple outputs. The picture that has to be drawn in such a case is multi dimensional and harder, if not impossible, to interpret, yet the measure of efficiency is still based on distance. This holds for the input oriented case as well.

2.2.2 From Allocative Efficiency to Profit and Cost Efficiency

So far, only the production function is used to measure efficiency. A producer however has not only to deal with a production function. Part of the profits and costs are determined by the prices of the inputs used and outputs produced. If prices are taken into account not every point on the production function in Figure 2.1 and 2.2 is efficient. Only points that maximize profit or minimize costs are fully efficient. For profit efficiency a producer has to maximize profits given the amount of input. The following example discusses how profit efficiency can be measured. A graphical representation is given in Figure 2.3.

The start of this example is the same as the output oriented TE example. Assume that one input ($X$) is available to produce two outputs ($Y_1$ and $Y_2$). The prices of these outputs are $p_1$ and $p_2$; the proportion of the prices is represented by line $A'A$. Because the price ratio is used to draw line $A'A$, every point on this line should generate the same amount of profit. The producer maximizes profit if line $A'A$ is shifted as far to the right as possible. This implies that a producer is profit efficient if she produces the amount where $A'A$ tangents the production curve $S'S$ (point $Q'$). Now suppose that the producer fails in setting the production to $Q'$ but produces $Q$. The producer is still technical efficient yet the allocation of the outputs
is inefficient. The allocation mismatch can be measured with the use of allocative efficiency (AE). For this measure a point on the $A'A$ line is needed that can be compared with the point $Q'$. Such a point is point $R$. This point has the same proportion of $Y_1$ and $Y_2$ as point $Q$ but is still located on the line $A'A$. The ratio of $OQ$ to $OR$ is now the measure of AE. The ratio has a value between zero and one. A value of one represents a fully allocative efficient producer. This can only be achieved if the producer produces on the point where the profit line is tangent to the production curve. This means that the producer chooses the right output mix. The measure gets the value of zero for the completely inefficient producer. This can only happen if the distance between the profit line and the production curve is infinite.

Now that TE and AE are discussed, they can be combined to measure profit efficiency (PE). Suppose that a producer generates outputs represented by point $P$. The above discussion shows that if the producer maximizes her profits she should produce on the $A'A$ line. The most suited point for evaluation is the point that has the same proportions of $Y_1$ and $Y_2$ as point $P$; this is point $R$. PE can be calculated by the ratio of $OP$ to $OR$. This measure will be equal to one if the producer is fully profit efficient and zero if no output is produced while she uses some inputs. Note that PE is a combination of TE and AE. If one knows the TE and AE and is interested in PE, one can just multiply TE and AE.

In the previous example it is shown how output oriented TE and AE determine PE. Following the same analogy input oriented TE and AE will determine cost efficiency CE. In the following example the input oriented case of

\textbf{Figure 2.3: Allocative and Profit Efficiency} \hspace{1cm} \textbf{Figure 2.4: Allocative and Cost Efficiency}
Chapter 2

AE will be shown. Figure 2.4 gives a graphical representation of this example. Suppose that a producer needs two inputs \( X_1 \) and \( X_2 \) with prices \( p_1 \) and \( p_2 \) to produce a certain amount of output \( Y \). Also in the example a line \( A'A \) is used that represents the proportion of input prices \( p_1 \) and \( p_2 \). To minimize the costs a producer has to shift this line as low as possible. If this is taken into account, the producer should set its inputs to the point where \( A'A \) is tangent to the production curve \( S'S \). This point is represented by point \( Q' \). Now suppose that a producer uses inputs represented by point \( Q \). Input AE can now be calculated by \( OQ \) over \( OR \) where \( R \) is a point that lies on the line \( A'A \) and has the same proportion of inputs as \( Q \).

Now that input oriented AE is discussed, it is straightforward to calculate CE. As shown above, PE is calculated by multiplying output oriented TE and AE. CE is calculated by multiplying input oriented TE and AE. Suppose that a producer uses inputs \( X_1 \) and \( X_2 \) as denoted by point \( P \). Its measure for CE will become the ratio of \( OP \) over \( OR \). Since \( OP \) is smaller or equal to \( OR \), the ratio is smaller or equal to one. A score of one is only achieved if the producer is fully efficient.

2.3 Estimating Efficiency

Now that the concept of efficiency has been discussed, it is useful to show how it can be estimated. In general, two types of efficiency estimators can be used. The first approach is data envelopment analysis (DEA) and the second is stochastic frontier analysis (SFA). The main difference between the two branches is that DEA is a non-parametric estimator that assumes a deterministic production function while SFA is a parametric and stochastic estimator. Without going into too much detail, DEA uses linear combinations of inputs and outputs of best practice producers to come up with an efficient frontier. The use of linear combinations implies that for a DEA model no functional form of the production function has to be specified. Furthermore a DEA model does not need any assumptions about the distribution of efficiency scores as is the case for SFA models as shown below. These nice features of DEA models come at the cost that DEA models do not take data errors and/or good and bad luck into account.\(^2\) If these are absent, DEA is a useful model to come

---

\(^2\) New types of models are developed that are of a hybrid form of DEA and SFA models (e.g. Kumbhakar et al., 2007). They have the nice properties of DEA while they allow for data errors.
up with efficiency estimates. Since this thesis especially focuses on the determining factors of efficiency rather than just efficiency scores, one of the functional properties of DEA cannot be exploited. Although it is possible to obtain efficiency scores without assuming a distribution on those efficiency scores, one has to choose a distribution if these efficiency scores are used in a regression with variables that try to explain the scores. Within the SFA framework estimators are developed that make it possible to include variables that determine efficiency simultaneously with the variables that determine the production function. Next to the property of including efficiency determining variables into the model, SFA also allows for data errors. These properties make SFA models suited to answer the questions raised in this thesis.\footnote{Although there is a recently developed DEA model that can deal with efficiency determining variables as well (Simar and Wilson, 2007), most of the papers on which the chapters of this thesis are based, were already in progress.} Since this thesis uses the SFA estimator extensively, the rest of this section shows how the concept of efficiency is implemented in this estimator. Because DEA is not used in this thesis, a discussion of DEA is beyond our scope. The interested reader is referred to Coelli et al. (2005) for an easy to grasp overview of DEA.

This section firstly describes how technical efficiency (TE) can be estimated with the use of SFA. Next it describes how SFA is used to estimate cost efficiency (CE). Subsequently, it describes how SFA can be used to estimate the effects that certain factors have on efficiency. Within this last part the model developed by Battese and Coelli (1993, 1995) is discussed. This model is the main model used in this thesis.

2.3.1 Stochastic Frontier Analysis and Technical efficiency

Suppose that output is completely determined by the inputs used via a production function that is the same for all producers in an industry. This can be written by Equation (2.1)

\[ y_i = f(x_i; \beta) \] (2.1)
In this equation $y_i$ is the output produced and $x_i$ are the $i$ inputs used by the producer or decision making unit (DMU) $i$. The function $f()$ represents the production function by the industry (e.g., a Cobb-Douglas or transcendental logarithmic production function). The $\beta$'s are equal for all the DMUs in the industry. The $\beta$'s can be analyzed by carrying out a regression of $y$ on the $x$'s. Since complete determination is assumed this regression should give a perfect fit. In this model all DMUs produce exactly the output that is predicted by the amount of inputs and the production function. If efficiency is included in the model, DMUs may produce less than the value predicted by Equation (2.1). This can be seen by multiplying the right hand side of Equation (2.1) by a parameter that has a value between zero and one.

$$y_i = f(x_i; \beta) \times TE_i$$

(2.2)

In Equation (2.2) the $TE_i$ term is the technical efficiency term DMU $i$ faces. If this efficiency term is zero the DMU will produce nothing irrespective of its production function and chosen quantities. The DMU is in this case completely inefficient. On the other hand, if the efficiency term is equal to one the DMU produces exactly the same as in Equation (2.1) and it is fully efficient. If the production function, the inputs and the output are known, the efficiency for each DMU can be determined by Equation (2.3).

$$TE_i = \frac{y_i}{f(x_i; \beta)}$$

(2.3)

No stochastics are involved so far, i.e., everything is completely deterministic. For real life applications this might be a bit too restrictive. If it is not noisy data that rejects complete determination it can be good or bad luck a DMU faces. Stochastics can be added into the model as is done in Equation (2.4)

$$y_i = f(x_i; \beta) \times \exp\{v_i\} \times TE_i$$

(2.4)

In this case $TE_i$ is determined by Equation (2.5)
Efficiency

\[ TE_i = \frac{y_i}{f(x_i; \beta) \cdot \exp(v_i)} \]  

So, technical efficiency depends on the random error \( \exp(v_i) \), meaning that the efficiency is stochastic. One should notice the difference with the model in Equation (2.3), where the efficiency is deterministic. However, this causes a circular reasoning. In order to estimate the technical efficiency component \( TE \), the production function with its parameters are needed and for estimating the parameters of the production function \( TE \) is needed. To break out of this circle some assumptions about the distribution of the noise term \( v \) and the efficiency component \( TE \) have to be made. First, suppose that \( TE_i = \exp(-u_i) \). Now the model, if transformed into log terms, gets the form of Equation (2.6) (if a Cobb-Douglas production function is assumed).

\[ \ln y_i = \beta_0 + \sum \beta_i \ln x_{i,n} + v_i - u_i \]  

The assumption made about the distribution of noise term \( v_i \) is, \( v_i \sim iidN(0, \sigma_v^2) \).

Since the efficiency term should have a negative influence on output it should be clear that \( u_i \) in Equation (2.6) should be positive. There is no consensus about the assumptions for the distribution of efficiency term \( u_i \). The inventors of SFA, Meesuen and Van den Broeck (1977) and Aigner et al. (1977), already discuss two types of distributions that can be used, i.e. an exponential or half normal distribution. Later extensions generalized these distributions to the gamma and truncated normal distribution with its truncation above zero. Yet, in general the assumption about the distribution of the efficiency term does not affect the estimated efficiencies (see, e.g., Cummins and Zi, 1998). To make sure that the model is properly identified and that noise really measures noise and not efficiency, it also has to be assumed that \( v_i \) and \( u_i \) are independent from each other.

If assumptions are made about the \( u_i \) and \( v_i \) terms, the model represented in Equation (2.6) can be estimated by using maximum likelihood (ML). The derivation
Chapter 2

of the likelihood functions is beyond the scope of this thesis, for an overview of the ML functions for the several distributions of the \( u_i \) terms I refer to Kumbhakar and Lovell (2000). Yet, the idea is to combine the \( v_i \) and \( u_i \) term in a combined error term \( \varepsilon_i \) term which result in a likelihood function. If this likelihood function is maximized, one can obtain the betas from Equation (2.6).

When the parameters of Equation (2.6) are determined the average of the combined error term \( \varepsilon \) can be calculated. Since the average of the noise term \( v \) should be zero it is clear that the average of the combined error term is equal to the average of the efficiency term \( u \). So the average of the \( \varepsilon_i \) is the average “efficiency” of the sample used. To calculate technical efficiency in line with the definitions given in part 2.2, one can use Equation (2.7).

\[
\overline{TE} = \exp\{-\overline{u}\}
\]  

(2.7)

Efficiency determination on DMU level was pioneered by Jondrow, et al. (1982). They developed an estimator that made it possible to estimate \( u \) conditional on \( \varepsilon \). Given the \( u \)’s, efficiency can be calculated with the use of Equation (2.8).

\[
TE_i = \exp\{-u_i^*\}
\]  

(2.8)

Here \( u^* \) are the obtained \( u \)'s of Equation (2.6) with the use of Jondrow et al. (1982).

2.3.2 Stochastic Frontier analysis and Cost Efficiency

The estimation of cost efficiency (CE) is similar to the estimation of technical efficiency (TE). Since CE is an input oriented approach, the decision making unit (DMU) faces a minimization problem. The SFA models that estimate CE differ from SFA models that estimate TE, apart from using different variables, in that an efficiency term is added rather than subtracted from the equation. The standard model is represented in Equation (2.9).

\[
c_i = c(y_i, w_i; \beta)\exp\{v_i + u_i\}
\]  

(2.9)

20
Efficiency

Where $c$ represents the total costs the DMU has. Function $c()$ is the cost function for the particular industry. If a Cobb-Douglas production function is assumed and the logs are taken from both sides of the equation, it can be rewritten as:

$$\ln c_i = \beta_0 + \beta_y \ln y + \sum_k \beta_k \ln w_{ik} + v_i + u_i$$  \hspace{1cm} (2.10)

The only difference between Equation (2.10) and Equation (2.6) is the positive sign before the $u_i$ term. The intuition behind the positive sign is that an inefficient producer faces higher cost than an efficient one. The assumptions for the cost efficiency models are the same as for the technical efficiency. The estimation of a cost efficiency model is, for this reason, similar to the technical efficiency model.

Cost efficiency can be calculated with the use of formula (2.11).

$$CE_i = \exp \{ u^*_i \}$$  \hspace{1cm} (2.11)

Where $u^*$ are the observed $u_i$’s of Equation (2.10). Although the $CE$ term in Equation (2.11) is called cost efficiency, it actually measures inefficiency. It has the value of 1 for a complete efficient producer and goes to infinity if a producer becomes more inefficient. This is in contrast with the technical efficiency term of Equation (2.8), which get the value of one for a complete efficient producer and a 0 for a completely inefficient producer. For this reason CE is also often calculated with Equation (2.8).

2.3.3 Determining Efficiency

So far it has been shown how efficiency can be estimated. Yet, an answer on what drives the efficiency can not be given with those models. In this section some models are discussed that deal with determining efficiency. This can give some insight in what factors drive efficiency. Two models are discussed. The first one is the two-stage method and the second one is the Battese and Coelli (1995) model. The advantages and disadvantages of both models are mentioned.
Two-stage Method

The application of the two stage method is simple. Firstly, the efficiency terms should be obtained by using one of the models discussed in earlier sections. Secondly, the obtained efficiency terms should be regressed on variables that are expected to influence efficiency.

\[ EFF_i = \delta_i + \sum \delta z_i + \varepsilon_i \]  \hspace{1cm} (2.12)

In Equation (2.9) \( EFF \) denotes technical or cost efficiency. The \( z_i \)'s are the variables that explain efficiency. This equation can be estimated with OLS or, since \( EFF \) lies between 0 and 1, with a Tobit model. The main advantage of this method is, as already mentioned, its simplicity. It has however, one major drawback. Most of the SFA models assume that efficiency is independent identical distributed while a second stage regression assumes that efficiency is dependent. Wang and Schmidt (2002) show that this violation of the assumption renders biased estimates that can become very severe. Although this method is still used every now and then, it is generally acknowledged to be inferior to the single-stage method.

Single-stage Method

To overcome the problem of biased estimates, variables that influence efficiency may already be included in the SFA model itself. In the 1990s several SFA models have been suggested to cope with variables that influence efficiency. Reifsneider and Stevenson (1991) extend the model by assuming that the efficiency term \( u_i \) from Equation (2.6) is composed by two factors.

\[ u_i = g(Z_i) + w_i \]  \hspace{1cm} (2.13)

\( Z_i \) in Equation (2.13) is a vector of firm specific efficiency explanatory variables. The term \( w_i \) captures unexplained efficiency with a distribution that is also assumed for the \( u \) term in Equation (2.6), i.e., this term is always positive. Furthermore it is assumed to have a positive range for function \( g() \). These assumptions cause that \( u_i \) is
also positive or at least zero just as in Equation (2.13). Function \( g() \) can be subtracted from the \( x \)'s from Equation (2.6). The ease of the approach suggested by Reifsneider and Stevenson (1991) is that the likelihood function of Equation (2.13) is equal to the one of Equation (2.6). One drawback however, is that the random errors around the efficiency term \( (w_i) \) can not become negative.

Kumbhakar et al. (1991) used equation (2.14) as a starting point to handle variables that influence efficiency.

\[
    u_i = \delta Z_i 
\]

(2.14)

In Equation (2.14) \( u_i \) is the efficiency term from Equation (2.6) and (2.10). \( Z_i \) is a vector of variables that are thought to influence efficiency. When deriving the log likelihood function, Kumbhakar et al. (1991) assume that \( u_i \) is normally distributed with a truncation above zero (Equation 2.15).

\[
    u_i \sim N^+ (\delta Z_i, \sigma_u^2) 
\]

(2.15)

Huang and Liu (1994) use Equation (2.14) as a starting point as well, but add an error term \( w_i \) that is assumed to be normally distributed, but is truncated (Equation 2.16).

\[
    w_i \leq -\delta Z_i 
\]

(2.16)

This restriction implies that \( u_i \) of Equation (2.14) is always equal to or larger than zero. With this restriction, it is possible to derive a likelihood function that can give an estimate for the betas of Equation (2.6), the deltas of Equation (2.16) and the variances of the noise term \( v \) and the efficiency term \( u \).

Battese and Coelli (1995) use the model of Huang and Liu (1994) and they show that the model can also be applied to panel data. Although Battese and Coelli do not change the model to exploit the panel properties, the model is often called a
panel SFA model. Although Huang and Liu (1994) were the first who developed this model, it is often referred to as the Battese and Coelli (1995) model.\footnote{A reason for this might be that Coelli (1996) developed a program often used to estimate SFA models and in particular this model. Since it is especially written for panel data he refers only to the paper of Battese and Coelli (1995).}

Although the previous model can handle panel data, it does not exploit the panel properties. Greene (2004) stresses the importance of exploiting panel properties since this makes it possible to differentiate between heterogeneity and efficiency. Heterogeneity in this sense is the difference between the cross-sections. It is possible that each cross-section has its own technology with a corresponding production function. If the data is pooled and efficiency is calculated, differences in technology can show up as differences in efficiency. Greene (2005, 2005a) discusses how fixed and random effects models can be implemented in the SFA framework. In Greene (2005a) he even shows how variables that influence efficiency can be included in a single stage model while incorporating the panel properties. His implementation uses an absolute normal distribution of the efficiency term where the first moment is modeled as in Equation (2.14). However, this implementation has one drawback: because of the use of an absolute normal distribution, the obtained signs for the parameters delta no longer have an economic interpretation.

In Chapter 3, 4, and 5 I use the Huang and Liu (1994) and the Battese and Coelli (1995) model. Since the model is applied on panel data the model from now on will just be called the Battese and Coelli (1995) model. The choice for this model is based on several criteria. Firstly, it can include variables that determine efficiency. Moreover, it can do this while it is still possible that the relationship between those variables and the efficiency component can have some error. Since this thesis focuses on the effects of institutions on efficiency this is a nice property. One difficulty that comes with this feature is that one cannot be certain whether a variable has a direct influence on output or whether this influence works via the efficiency term. Chapter 7 deals with this problem. Secondly, the Battese and Coelli (1995) model is often adapted in the bank efficiency literature. This makes it easier to compare our results with these of other studies. Because SFA models give relative efficiency scores, the efficiency score of bank A obtained from study 1
cannot be compared with the efficiency score of bank B obtained from study 2. Yet, it is possible to compare the parameters of delta. One major problem of the Battese and Coelli (1995) model is that it does not exploit panel properties. To overcome this shortcoming country dummies are sometimes added as suggested by Greene (2004). In the ideal case bank dummies should be added but because of the large amount of cross-sections with respect to the time dimension this is infeasible.

2.4 Estimating the Efficiency of Banks

In the previous sections it has been shown how efficiency can be determined. In this section I show how this approach can be applied to the banking sector and what choices one has to make. For estimating efficiency, no matter whether it is technical efficiency, cost efficiency, or profit efficiency, measures of inputs and outputs are needed. However, for banks it is not directly clear what inputs and outputs are. The determination of inputs and outputs depend on the role of a bank and there is a debate in the literature to appoint this role (Colwell and Davis, 1992). Theory came up with two competing models. In the first model, the production model, a bank is seen as a firm that uses capital and labor to produce deposits and loan accounts (Benston, 1965). In this model outputs are determined by the number of deposits and loan accounts and the costs in this model are operating costs. In the second model, the intermediation model, a bank is treated as an intermediary who uses deposits to generate loans (Sealy and Lindley, 1977). Within this model labor and capital are the inputs used to produce loans and thus the costs in this model are determined by operating costs and interest costs.

Each of these models has its strengths and its weaknesses. One of the weaknesses of the intermediation model is that it uses stock variables rather than flow variables for its inputs and outputs. It is a natural way to think of inputs and outputs as a certain amount used per time period. However, with stock variables one looks at the total amount at a certain point in time. The production model solves this problem by using flow variables, yet, it comes at the cost that it is difficult to weight the different types of accounts to their contribution on total output (Benston et al., 1982). The weakness of the intermediation model of using stock variables is also its strength. There are only few datasets available that have information about flow
variables like the number of accounts. Miller and Noulas (1996) already state that the choice of the model depends on the availability of the data. Since this thesis needs a dataset that covers multiple countries and there is no dataset that covers multiple countries and has flow variables, the intermediation model is the model to use. Besides this pragmatic reason there is also a theoretical reason why an intermediation model is chosen. In Chapter 6 the link between an efficient financial sector and economic growth will be examined. The role of the financial sector in this context is an intermediation role, therefore, the intermediation model is more appropriate.

After the inputs and the outputs are specified, one can choose the type of efficiency that will be estimated. If only data on inputs and outputs are available, the choice is simple. Because one needs data on input or output prices for cost (CE) or profit efficiency (PE), only technical efficiency (TE) can be estimated. However, if there are several outputs, it is difficult to estimate TE. In the SFA framework discussed above, only one output can be specified. Although this problem can be overcome by estimating a system of equations, this procedure is cumbersome. If there are multiple outputs and data on input or output prices are not available, it might be the best idea to use a DEA type of estimator, since DEA can deal with multiple in- and outputs. If data on input or output prices are available and there are multiple outputs it is possible to estimate cost or profit efficiency by using the duality property of production functions and transform it to a cost or profit function. Although it is possible to estimate cost and profit efficiency, each choice comes with some assumptions. To estimate CE one has to assume that the bank is a price taker for its input prices. If the bank has some market power to set input prices, these input prices are no longer exogenous but vary with the demand for inputs and thus with the input mix. This makes it theoretically possible that by choosing another input mix prices change in such a way that costs decrease. This effect makes the cost efficiency measure that is explained in Section 2.2.2 useless. Profit efficiency adds to this that output prices should also be exogenous. Adams et al. (2002) show that there is some market power in the banking sector on the demand side as well as on the supply side. However, they also show that assuming no market power on one side does not affect results. This outcome might be a reason
to favor analyzing cost efficiency for the banking sector. There is another, more technical, reason why estimating cost efficiency might be preferred. If one is estimating a cost and a profit function, the fit of the cost function is in general much better. This means that the error term \( v \) in Equations (2.6) and (2.10) is relatively low compared to the efficiency term \( u \). Because of the better fit it is easier to estimate a correct efficiency term. Because of these reasons this thesis focuses on cost efficiency.

After the inputs and outputs are specified, one has to choose the type of the production function. In Section 2.3 a standard Cobb-Douglas (CD) production function was used. Although this type of production function is often used in economics it is arguably to restrictive for efficiency analysis (Coelli et al., 2005). Because of the restrictive nature of the CD production function, the efficiency of banks can be underestimated. This means that banks operate closer to the real frontier, i.e., the best practice production function, than what is found, just because the production function is not flexible enough. To overcome this problem more flexible production functions are assumed. One of the most often used production functions of this type is the transcendental logarithmic (or translog for short) production function (Christensen et al., 1973). This production function includes (after a log transformation) not only the separate inputs but also an interaction of the inputs and the inputs squared. The CD production function is a special case of the translog production function. So, if a CD model describes the data well, the coefficients of the interactions and inputs squared will all be zero. Although the translog production function is already quite flexible, some authors argue that the production function should be even more flexible (see e.g. Berger and Mester, 1997). The type of production functions that are used to achieve this are the so called Fourier type production functions. They use a Fourier transformation of the translog function. With a Fourier transformation sines and cosines are included into the production function. However, one drawback of this type of functions is that they do not have an economic interpretation anymore. While the interpretation of the translog production function is already difficult, the economic interpretation of sines and cosines within production theory is impossible. Within this thesis the translog form is used. It adds flexibility while still being interpretable.
Chapter 2

2.5 Conclusion

Within this chapter the theoretical foundations are laid that are used in the rest of this thesis. The chapter discussed the type of efficiencies and how they can be estimated. It ended with a discussion on how the theory can be used to estimate bank efficiency. This section, however, only discussed the models that are used in this thesis and didn’t touch on the recent developments in efficiency estimators. It is a field that is under development. Although there are some books that treat the subject very well (e.g. Kumbhakar and Lovell, 2000 and Coelli et al., 2005) they miss the latest developments. Journals in which papers on this issue are published occasionally include the European Journal of Operational Research, Journal of Econometrics, and Journal of Productivity Analysis.