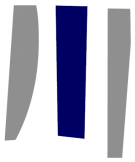


European Investment Bank

Economic and Financial Report 1999/02

The Recursive Thick Frontier Approach to Estimating Efficiency

Rien Wagenvoort and Paul Schure

European Investment Bank

100, blvd. Konrad Adenauer

L-2950 Luxembourg

FAX: (352) 4379-3492
Email: infoefs@eib.org

Notes

Rien Wagenvoort is an Economist at the EIB. Paul Schure is a Ph.D. candidate in Economics at the European University Institute in Florence. Part of this research was done while the second author visited the EIB.

ECONOMIC AND FINANCIAL REPORTS are preliminary material circulated to stimulate discussion and critical comment. Quotation of material in the *Reports* should be cleared with the author or authors.

The views expressed are those of the individual authors, and do not necessarily reflect the position of the EIB.

Individual copies of the *Reports* may be obtained free of charge by writing to the above address or on-line from http://www.eib.org/efs/pubs.htm

Non-Technical Summary

The technology of a firm determines the maximum possible output that can be produced given a bundle of inputs and can be represented by the production function. Once the functional form and the parameters of this function are known, one may evaluate the technological and managerial efficiency of a company by measuring the distance between its realised production and the relevant position on the production curve. Under weak conditions the principle of duality applies and thus the production function is directly related to the cost function. A cost function relates the minimum cost incurred for producing a certain mix and level of outputs given the input prices. The purpose of this note is to provide a regression technique for estimating the unknown parameters of the production or cost function when pooled cross-sections and time series, so-called panel data, are available.

Estimating technology functions requires non-standard regression techniques. The reason for this is that we look at the minimum costs incurred (or the maximum output produced) instead of the average costs (average output). In this paper we adopt an estimation method which takes into account that deviations from the cost or production function, the so-called frontier, may emerge due to inefficiency but also due to other temporary firm specific reasons (for example, re-organisation costs) or simply bad and good luck or measurement errors. Distinguishing between *randomness* and *efficiency* however is not trivial. *Stochastic* or *thick* frontier approaches[1] have to impose assumptions that determine how random effects can be separated from other (wasting) effects. The main problem with the traditional methods is that these assumptions are, whether feasible or not, difficult to test. This is a serious weakness because the distinction between randomness and inefficiency remains in this way somewhat arbitrary.

The method that we propose, the Recursive Thick Frontier Approach (RTFA), is less vulnerable to the criticism mentioned above. Instead of making the usual distributional assumptions when applying the Stochastic Frontier Approach (SFA) we assume that the probability of an efficient firm of being at either side of the cost frontier is equal to one half. This assumption can be tested for a selected sample of

[1] The frontier is called *stochastic* or *thick* in order to indicate that best practice firms are allowed to be positioned close to the frontier but not necessarily at the frontier.

best practice or so-called X-efficient companies. We therefore consider a selection criterion that sorts the sample into a group of X-efficient and a group of X-*in*efficient companies. The cost frontier is estimated using only the observations of the former group. If our test statistic rejects that on average the probability for a firm to be above or below the regression line is ½, then we reduce this group of firms by eliminating those companies which are relatively far positioned above the regression line (i.e. with relatively high costs) in the case one estimates a cost function. Our method is only suitable for panel data. The time dimension of panel data enables to require information on the persistence of some firms of having lower cost than others, and this is obviously not available in analyses of single cross-sections. Therefore, we argue that it will be always difficult to distinguish between luck or efficiency if only single cross-sections are used to estimate the frontier.

To investigate the performance of RTFA, we simulate a panel data model where half of the observations are drawn from best practice firms and the other half from X-inefficient firms. Although the data seems to perfectly suit SFA, our results show that RTFA produces considerably more reliable parameter estimates.

Abstract

The traditional econometric techniques for frontier models, namely the Stochastic Frontier Approach (SFA), the Thick Frontier Approach (TFA) and the Distribution Free Approach (DFA) have in common that they depend on *a priori* assumptions that are, whether feasible or not, difficult to test. This paper introduces the Recursive Thick Frontier Approach (RTFA) to the estimation of technology parameters when panel data is available. Our approach is based on the assertion that if deviations from the frontier of X-efficient companies are completely random then one must observe for this group of firms that the probability of being located either above or below the frontier is equal to one half. This hypothesis can be tested for panel data sets but requires sorting of the full sample into a group of X-*in*efficient firms and a group of X-efficient (best practice) firms. The cost frontier is estimated using only the observations of the latter category.

JEL Classification Numbers: C13, C23, C40

Keywords: Cost/Production Function, Thick Frontier Approach, X-efficiency

1. The Problem: Choosing Between Randomness and Differences in Efficiency

The technology of a firm determines the maximum possible output that can be produced given a bundle of inputs and can be represented by the production function. Once the functional form and the parameters of this function are known, one may evaluate the technological and managerial efficiency of a company by measuring the distance between its realised production and the relevant position on the production curve. Under weak conditions the principle of duality applies and thus the production function is directly related to the cost function. A cost function relates the minimum cost incurred for producing a certain mix and level of outputs given the input prices. The purpose of this note is to provide a regression technique for estimating the unknown parameters of the production or cost function when pooled cross-sections and time series, so-called panel data, are available.

If one discerns that measurement errors in the output variables may arise or that production may fluctuate due to factors which are beyond the control of the firm's management then one must allow that part of the differences in the measured output among firms are caused by random effects rather than differences in management efficiency or technology. The econometric approaches to estimating frontier models make this distinction between *randomness* and *efficiency* in contrast with the mathematical approach where, roughly speaking, any deviation from the frontier is assumed to reflect inefficiency. The Stochastic Frontier Approach (SFA), the Distribution Free Approach (DFA), and the Thick Frontier Approach (TFA) belong to the group of econometric techniques[2] while Data Envelopment Analyses (DEA) is the common name for the mathematical programming approach.

When applying the econometric methods we need to adopt, one way or the other, a rule (or impose an assumption) that determines how random effects can be separated from other (wasting) effects. The main problem with the traditional econometric frontier methods is that they depend on *a priori* assumptions that are, whether feasible or not, difficult to test. This is a serious weakness because the distinction between randomness and inefficiency remains in this way somewhat arbitrary.

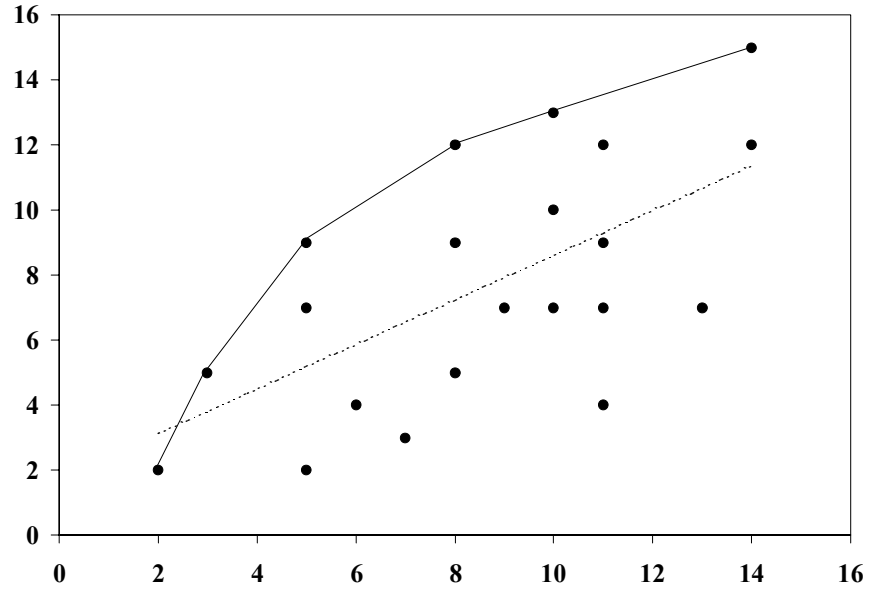
In this paper we develop the Recursive Thick Frontier Approach (RTFA). Our method is based on the assertion that if deviations from the frontier of X-efficient, i.e. best practice, companies are completely random then one must observe for this group of firms that the probability of being located either above or below the frontier is equal to one half.[3] This hypothesis can be tested for panel data sets but requires sorting of the full sample into a group of X-*in*efficient firms and a group of X-efficient firms. The cost frontier is estimated using only the observations of the latter category.

In the following section we review the commonly applied approaches to estimating the production or cost frontier before defining our panel data model in section 3 and introducing RTFA in section 4. In section 5 we discuss under which circumstances RTFA may go wrong and how to prevent such a situation. Section 6 contains an

[2] See Aigner, Lovell and Schmidt (1977), Berger (1993) and Berger and Humphrey (1992) respectively.

[3] A similar assumption is made when applying the Distribution Free Approach. In this case, the random effects to the production of a specific firm are assumed to cancel each other out over time. However, Berger and Humphrey (1992) additionally assume that inefficiency is persistent over time. The latter rather strong assumption does not have to hold for successful application of the RTFA.

Figure 1 Two extreme choices for the production function (DEA versus OLS)[a]



[a]See Figure 1-1, Charnes, Cooper, Lewin and Seiford (1994).

illustrative example where the performance of both the traditional SFA and RTFA in a simulated panel data model is evaluated. Finally, section 7 concludes.

2. A Brief Overview of Frontier Analysis

2.1 Data Envelopment Analyses (DEA)

Figure 1 plots the observations on input-output combinations of some firms. Mathematical programming techniques can be employed in order to find the close fitting frontier which envelops all data points (see the solid line in Figure 1). In this case, the production function is completely determined by the most "efficient" companies in the sample. In Figure 1, the frontier is allowed to be discrete and piecewise linear. Evidently, other functional forms can be chosen. Data envelopment analysis was introduced by the pioneers Charnes, Cooper and Rhodes in 1978.[4]

Note however that the production function is constructed on the basis of the information contained in the data. Therefore, only relative efficiency measures are considered. According to our definition given in the previous section, however, the production function is defined by an absolute efficiency criterion. We thus have to make the implicit assumption that some of the firms in the data set are efficiently operating.

2.2 Stochastic Frontier Approach (SFA)

[4] Finding parametric production functions by using mathematical programming techniques already began with the work of Aigner and Chu in 1968.

Two main criticisms of using mathematical programming techniques are mentioned in the existing literature on frontier analysis. First, the methods are extremely sensitive to outlying observations. Only one observation may cause a shift in the frontier. This observation however might emerge from a measurement error and as a consequence might overstate the technological capacity of the industry. Second, any observation which lies below the frontier is marked as a relatively inefficient company. Such an indication only makes sense under the assumption that the management of the firm has perfect control over any factor that may affect total output. If, on the contrary, there are measurement errors, unobservable shocks or factors which are beyond the sphere of influence of the management, then we must allow companies to fluctuate around the frontier without necessarily being inefficient. A firm can be efficient in the sense that its management makes rational and optimal decisions under uncertainty but at the same time, after revelation of the state of nature, it may not be positioned on the production function. An extreme opinion would be to say that all deviations from the frontier are due to bad or good luck and measurement errors. In this case, the best course we can follow to estimate the production function is to apply a standard regression technique such as OLS (see the dashed line in Figure 1).

In response to these arguments, Aigner, Lovell and Schmidt (1977) develop a stochastic frontier production model by appending a random disturbance term to the production function. The error term is assumed to be the sum of two random components, a noise term which is symmetrically distributed around zero (to model measurement errors and unobservable shocks) and an error component which is strictly negative (to measure inefficiency). The model can be written as

$$y_i = f(x_i;\beta) + u_i + v_i \qquad (1)$$

where y_i is the output, $f(x_i;\beta)$ is the production function with unknown parameter vector β, v_i represents the symmetric disturbance and u_i determines the inefficiency of company i ($u_i \leq 0$). Usually the production function is chosen to be log linear in its arguments, v_i is assumed to be independently and identically distributed (iid) following the normal distribution and independently generated from the inefficiency terms u_i which, in turn, are also iid and assumed to follow, for instance, a truncated normal or an exponential distribution. Maximum likelihood procedures are used to estimate the unknown parameters of model (1). Jondrow, Lovell, Materov and Schmidt (1982) show how to disentangle the inefficient component from the entire error term $\varepsilon_i = u_i + v_i$ by considering the expected value of u_i conditional on ε_i.

Although the Stochastic Frontier Approach (SFA) provides a solution to the second criticism mentioned above it is still vulnerable to outlying observations. The classical Maximum Likelihood estimators have a breakdown point of zero, i.e. only one observation may cause the estimator to return any outcome.[5] In this paper we will consider a regression technique which is less sensitive to outliers.

Another argument that was brought up to criticise the SFA is the arbitrary choice of the distributional assumption concerning the inefficiency component of the error term.

[5] The maximum fraction of data contamination which leaves the estimator determinate defines its breakdown point (see Donoho and Huber (1983)).

Not so much this assumption in itself as the lack of a testing procedure makes SFA questionable.[6] Moreover, in a panel data framework, it may occur that the SFA regression residuals are approximately normally distributed for each single cross-section.[7] In this case, the researcher would be tempted to conclude that the one-sided error component is negligible and therefore the companies in the sample are equally efficient. However, when comparing the separate cross-section results, one may find that the observations of many companies are at the same tail side of the normal distribution for each year of the sample period. If the time horizon is long enough, then it seems wrong to ascribe this result to bad and good luck. In other words, while not being revealed by SFA, there are substantial differences in production efficiency. Some companies have persistently higher production than others. This does not necessarily induce asymmetric errors in the frontier model. Although for a single cross-section we cannot do more than accepting the SFA efficiency hypothesis, SFA does clearly not provide an appropriate procedure for distinguishing best practice from inefficient firms if the full panel data set is to be taken into account.

2.3 Distribution Free Approach (DFA)

When both cross-sections and time-series are available then solutions to circumvent restrictive distributional assumptions are at hand. For instance, Berger (1993) calls his method "distribution free" since no specific distribution for the inefficiency component u_i is chosen. However, Berger assumes that managerial inefficiency is persistent and constant over time and thus in a panel data context one can write $u_{ti} = u_i$. On the other hand, the random error v_{ti} will cancel out over the years. DFA involves estimation of the panel data model:

$$\ln TC_{ti} = \ln C_t(Y_{it}, w_{it}) + \ln u_i + \ln v_{ti} \qquad (2)$$

where TC is the total costs of firm i in period t, C_t is the industry cost function in period t, Y_{ti} is the output vector and w_{ti} is a vector of input prices and ln represents the natural logarithm. Zellner's Seemingly Unrelated Regression (SUR) estimator is used to estimate model (2) with composite disturbance $\varepsilon_{ti} = \ln u_i + \ln v_{ti}$. The average of the regression residuals per cross-sectional unit i is then computed to estimate $\ln u_i$.

The following conditions must hold to successfully apply DFA: $u_i \in [1, \infty)$, $E[\ln v_{ti}] = 0$ and the usual orthogonality condition must be satisfied. If the cost function contains a constant then no unbiased estimate of the inefficiency component $\ln u_i$ can be obtained. However, the relative X-efficiency measure:

---

[6] Kopp and Mullahy (1990) introduce a Generalized Method of Moments (GMM) estimation procedure for frontier models which enables various degrees of distributional flexibility and provides moment-based specification tests. Rather than imposing a arbitrarily chosen distribution for the inefficiency component u_i, a parametric relationship between the first and third moments of u_i need to be specified. This specification is *ad hoc* in itself but an important aspect of Kopp and Mullahy's procedure is that it enables to test the validness of the distribution of the one-sided error component.

[7] This result was found in a study of the cost efficiency of almost 2000 European banks by Schure and Wagenvoort (1999).

$$XEFF_i = \exp(\ln u^*_{\min} - \ln u^*_i) = \frac{u^*_{\min}}{u^*_i} \qquad (3)$$

is still accurate in this case. $\ln u^*_{\min}$ is the minimum of $\ln u^*_i$ where the latter is the estimate of $\ln u_i$. X-efficiency refers to a measure of managerial/operational efficiency and can be contrasted with scope or scale efficiencies. The measure $XEFF_i$ is equal to 1 for an efficient firm and takes lower values otherwise.

2.4 Thick Frontier Approach (TFA)

Although DFA is less dependent on *a priori* distributional assumptions than SFA, it relies on the strong assumption that X-efficiencies are constant over time. If there are changes in X-efficiency, then one can only predict the average inefficiency over the past for a certain firm.

Berger and Humphrey (1992) consider another distribution free way of estimating cost frontiers using panel data, the so-called "Thick Frontier" Approach (TFA). This method starts with sorting of the data on the average costs.[8] It proceeds with the estimation of two "thick-frontiers", one for the lowest and one for the highest average costs quartile of firms. These regressions are independently executed for each year in the sample. Average inefficiency of the highest quartile companies is then computed by comparison of the two thick frontiers (see Berger and Humphrey (1992) for details). Even if the errors associated with those separate cost functions are not drawn from a random variable which is symmetrically distributed around zero, i.e. the lowest quartile may still contain some inefficient firms (not only randomness) then TFA may still provide a useful comparison of high and low cost firms. On the other hand, only in rare cases the actual production frontier can be found in such a way. As a consequence, results regarding the average production inefficiency within the sector will usually be biased downward, i.e. efficiency will be overstated.

In this paper we consider a TFA type of regression technique. The main problem of the frontier approaches mentioned so far is that the choice between random error or inefficiency remains somewhat arbitrary: DEA ignores randomness from the very beginning, SFA results depend on *a priori* distributional assumptions, DFA makes strong assumptions on the evolution of X-efficiency over time and last, TFA sorts the data in arbitrarily selected groups of firms, i.e. instead of quartiles other quantiles can be chosen. Therefore we suggest to apply a formal test in order to reveal whether the frontier lies in a cloud of observations which belong to relatively efficient firms. With respect to this group of companies any deviation from the frontier must be random. If the test statistic, i.e. the Lagrange Multiplier (LM) test of Breusch and Pagan (1980) or a "Binomial test" statistic, rejects randomness of the error terms then a smaller quantile is chosen and a new frontier is computed. The algorithm is called Recursive Thick Frontier Approach (RTFA) and uses the Trimmed Least Squares (TLS) estimator (see Koenker and Bassett (1978)).

In the subsequent section we discuss the assumptions of our panel data model and argue why these conditions bring forward to consider Breusch and Pagan (1980) LM test or a "Binomial test".

[8] Berger and Humphrey (1992) define average costs as total costs divided by total assets.

3. Assumptions of the Model

Suppose there are n cross-sectional units $(i = 1,..,n)$ and T time periods $(t = 1,..,T)$. Thus the full sample contains nT observations. Consider the linear panel data model

$$y_{ti} = x_{ti}\beta + \varepsilon_{ti},\ i \in E \qquad (4)$$

which describes the relationship between output y_{ti} and a k-dimensional input bundle x_{ti}.[9] As usual, β is a k-dimensional column vector of unknown parameters and ε_{ti} is the error term associated with firm i in period t. Note that equation (4) is only expected to hold for the most efficient companies in the sample, i.e. for companies with subscripts i which belong to the set E. Not all n companies are necessarily in E. Here efficiency is defined in terms of X-efficiency rather than scale or scope efficiency. Thus, given the output mix and level of inputs the management of the firm optimally allocates the firm's resources.

If there are companies which are included in the integer set E and at the same time are persistently located above or below the production curve defined by relationship (4) then the set E has not been well defined. In this case, relatively efficient firms are then the firms with positive errors in all periods in comparison to relatively inefficient firms with concomitant negative errors. For the right choice of E, however, observations on the same cross-sectional unit are expected to be randomly distributed around the frontier. Evidently, especially for panel data with short time series, this does not imply that the regression residuals of a certain company can be incidentally of the same sign. The probability of finding a positive or negative residual however must be equal. To conclude, by definition, the disturbances of model (4) are random and do not reflect managerial inefficiencies.

We assume that the following conditions must hold for panel data model (4):

Assumptions:

(A.1) ε_{ti} are independent and identically distributed with symmetric distribution function F around zero, $E[\varepsilon_{ti}] = 0$, for $i \in E$.

(A.2) The covariance matrix of the joint disturbance vector ε, $E[\varepsilon\varepsilon'] = \Omega$, is diagonal, i.e. $E[\varepsilon_{ti}\varepsilon_{si}] = 0$ for $t \neq s$, $i \in E$, and $E[\varepsilon_{ti}\varepsilon_{tj}] = 0$ for $i \neq j$, $i, j \in E$.

(A.3) The orthogonality condition $E[\varepsilon_{ti}x_{ti}] = 0$ holds for $i = 1,..,n$.

Note that assumptions A.1-A.2 do not hold for the errors (defined with respect to relationship (4)) associated with firms which do not belong to the set E, i.e. the relatively inefficient firms.

In section 4 the algorithm RTFA is presented and a X-efficiency measure is discussed.

[9] Our exposition follows the estimation of a production function. Evidently, the proposed method can be used for the estimation of a cost function.

4. A Recursive Algorithm for the Computation of the Stochastic Frontier in Panel Data Models

A procedure for testing the diagonality of the covariance matrix Ω (assumption A.2) was proposed, among others, by Breusch and Pagan (1980), assuming normality of the disturbances in equation (4). They show that, under the null hypothesis of zero autocorrelation (more precisely, under assumption (A.2)), the Lagrange Multiplier statistic

$$\lambda_{LM} = n\sum_{t=2}^{T}\sum_{s=1}^{t-1} r_{ts}^2 \qquad (5)$$

has asymptotically a χ^2-distribution with $T(T-1)/2$ degrees of freedom, where the correlation coefficient

$$r_{ts} = \frac{\omega_{ts}^*}{\sqrt{\omega_{tt}^*\omega_{ss}^*}} \qquad (6)$$

is computed with the estimated covariances $\omega_{ts}^* = \frac{1}{n}\sum_{i=1}^{n}(y_{ti} - x_{ti}\beta^*)(y_{si} - x_{si}\beta^*)$ and β^* is an estimate of β.

RTFA begins with an Ordinary Least Squares regression using the full sample of observations. On the 1% significance level, the computed LM statistic indicates whether all companies in the data set can be considered as equally efficient. If not, then we reduce the set E. In practice, δ% of the firms with the lowest mean (over time) of the residuals are left out from the sample. Then we repeat the regression and computation of the LM statistic for the reduced sample until assumption A.2 cannot be rejected, i.e. the largest possible group of relatively efficient firms has been identified. Details of the algorithm are summarised in the annex.

Although RTFA will eliminate step by step the relatively inefficient companies, including the outlying observations which lie far below the production frontier, outliers which are positioned above the production frontier may still push the regression line too far up. In order to obtain outlier robust estimates we employ an *one-sided* Trimmed Least Squares (TLS) estimator to estimate β.[10] Those observations with concomitant value of the standardised regression residual, $(y_{ti} - x_{ti}\beta^*)/\sigma^*$, that is lower than the first percentile of the standard normal distribution (-2.54) are left out after the initial OLS regression. A robust estimate, σ^*, of the standard deviation of the regression residuals is computed with the help of the Median Absolute Deviation (MAD) estimator:[11]

$$MAD(r) = med(|r - med(r)|)/0.6745 \qquad (7)$$

Two remarks are called for. First, note that we select the efficient firms on the basis of their distance to the regression line instead of their average costs as was suggested by

[10] Strictly speaking, the TLS estimator does not possess the desired robustness properties of having a good breakdown point or a bounded influence function. We therefore recommend to use other outlier robust estimators such as a HBP GM technique (see for instance, Hinloopen and Wagenvoort (1997)) instead of the TLS. Due to restrictions on available computing time we do not apply those more computing intensive estimators.

[11] See, among others, Rousseeuw and Leroy (1987), p.45.

Berger and Humphrey (1992). The latter approach will omit relatively small companies if there are increasing returns to scale, even when they efficiently allocate their resources. Since we are primarily interested in X-efficiencies, our selection criterion more appropriately sorts the data. Second, even in the case where the observations of both the inefficient and efficient companies are drawn from a normal distribution it is unlikely that the computed residuals of the regression equation are exactly normally distributed. This can be imputed to the fact that when discarding observations corresponding to inefficient companies, some data points associated with efficient firms may be discarded as well (due to *bad luck* X-efficient companies). That is, after convergence of the RTFA algorithm, the distribution of the RTFA residuals will be truncated from below.[12] The LM statistic requires normality of the regression residuals. Therefore, we also consider another test which is less dependent on the distribution of the regression residuals.

Another way of formulating the assumption that efficient companies are randomly distributed around the frontier is to assume that the conditional probability of being positioned either above $(\varepsilon_{ti} > 0)$ or below $(\varepsilon_{ti} < 0)$ the frontier is equal to 0.5 given any value for the lagged ε_{ti} instead of assuming (A.2):

Assumption: (A.4) $\Pr(\varepsilon_{ti} > 0 | \varepsilon_{si}, s = 1,..,t-1) = \Pr(\varepsilon_{ti} < 0 | \varepsilon_{si}, s = 1,..,t-1) = 0.5$.

Now we define the following random variable

$$Z = \sum_{i=1}^{n} indic_i \,, \qquad (8)$$

where $indic_i = 1$ if the event "$T-1$ or T of the residuals r_{ti} are positive" occurs or the event "$T-1$ or T of the residuals r_{ti} are negative" occurs, $indic_i = 0$ otherwise. The random variable Z has a binomial distribution with probability p that the indicator function $indic_i$ returns 1.[13] For large samples (in n) and when probability p is not too small, the binomial distribution approximates to the normal distribution. Therefore, we suggest to compute the following "Binomial test" statistic

$$\lambda_B = \frac{(Z - np)^2}{np(1-p)} \qquad (9)$$

in order to test assumption (A.4). λ_B is asymptotically chi-squared distributed with one degree of freedom.

Once the frontier is established, X-efficiency is computed as:

$$XEFF_{ti} = y_{ti} / x_{ti}\beta^*_{TLS}, \qquad (10a)$$

in the case of the production approach. In the case of the cost function approach, the numerator and denumerator of formula (10a) are swapped:

[12] Strictly speaking this distribution will also be truncated from above since we apply the TLS estimator.

[13] For $T = 5$, $p = 12 * 0.5^5 = 0.375$.

$$XEFF_{ti} = x_{ti}\beta^{*}_{TLS} / y_{ti}, \qquad (10b)$$

To reduce the effect of randomness, average values per cross-sectional unit (or per time period) can be calculated.

5. Pitfalls

Two pitfalls however have to be taken into account. First, when technological progress is relevant the Breusch and Pagan (1980) Lagrange Multiplier test and the Binomial test will reject assumption A.2 and A.4 respectively, even for the sample of relatively efficient firms. In this case, observations on those companies will lie below the production frontier as defined in (4) for early periods and above the production curve for later episodes. This problem can be solved by the introduction of time dummies to measure technological improvement or by introducing for each time period t a different parameter vector β_t in model (4). The second pitfall concerns the adjustment for outlying observations. If there are only a few efficient companies in the full sample then it may happen that RTFA detects a frontier which is still relatively far positioned from these outlying observations. As a consequence, the researcher must further investigate the nature of the reported outliers before ascertaining the industry production technology. Note however that outlier robust regression techniques are indispensable for the detection of these anomalous observations in the first place. If the number of observations around the frontier is sufficiently high then RTFA provides correct estimates of the technology parameters even in the presence of outliers.[14]

6. An Example: Estimation of the Production Function

This section contains a simulation experiment. Figure 2 is a typical diagram of a pool of efficient (crosses) and less efficient (triangles) firms among 500 competitors in the industry over five years. The picture shows, at least for the crosses, a linear positive relationship between output and input.

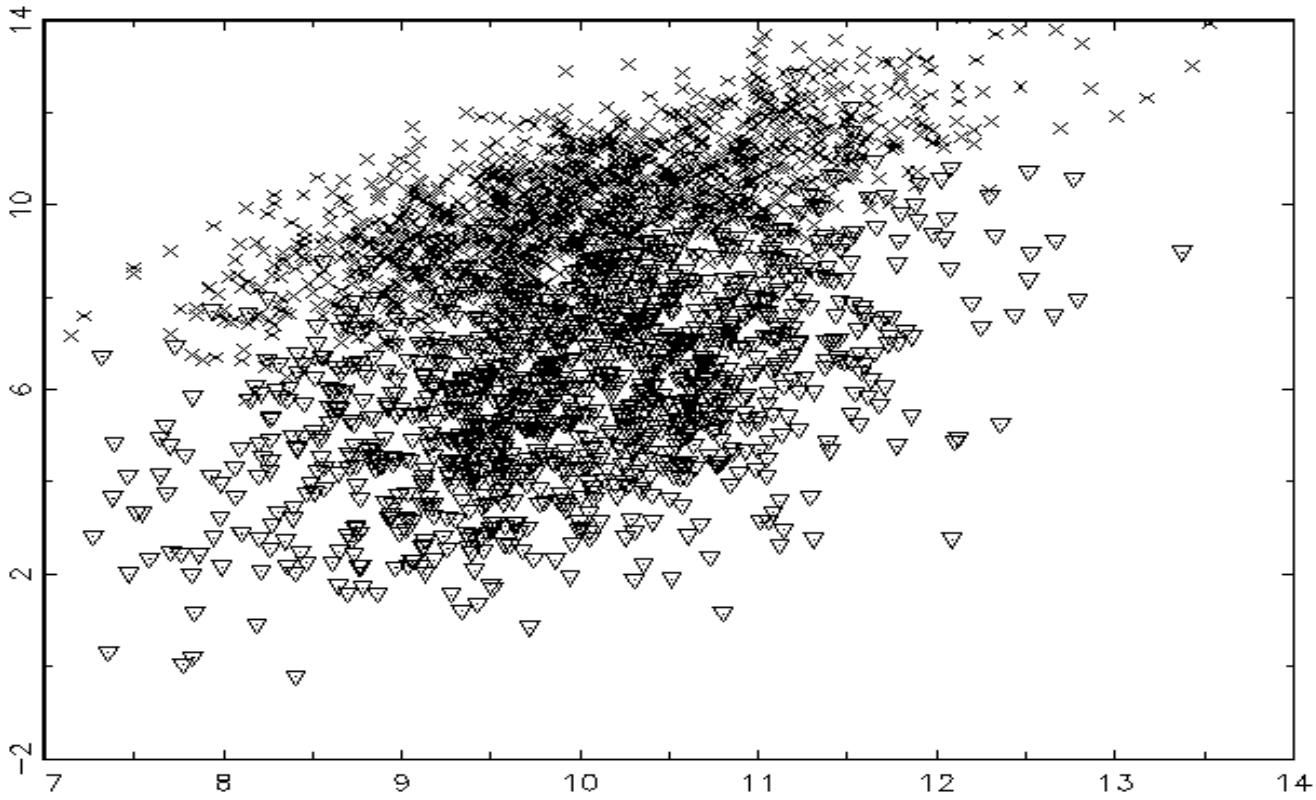
The data of Figure 2 was generated by the following Data Generating Process (DGP):

$$\begin{aligned}
x_{ti} &= 10 + \eta_{ti}, \;\; i = 1,..,500, \;\; t = 1,..,5 \\
\varepsilon_{ti} &= -5 - u_i + v_{ti}, \;\; i = 1,..,250, \;\; t = 1 \\
\varepsilon_{ti} &= -4 - u_i + v_{ti}, \;\; i = 1,..,250, \;\; t = 2 \\
\varepsilon_{ti} &= -3 - u_i + v_{ti}, \;\; i = 1,..,250, \;\; t = 3 \\
\varepsilon_{ti} &= -2 - u_i + v_{ti}, \;\; i = 1,..,250, \;\; t = 4 \\
\varepsilon_{ti} &= -1 - u_i + v_{ti}, \;\; i = 1,..,250, \;\; t = 5 \\
\varepsilon_{ti} &= v_{ti}, \;\; i = 251,..,500, \;\; t = 1,..,5 \\
y_{ti} &= x_{ti} + \varepsilon_{ti}
\end{aligned} \qquad (11)$$

where η_{ti} and v_{ti} are independent and standard normally distributed and u_i is drawn from a standard half normal distribution. Note that the set of efficient firms

[14] The number of best practice firms must exceed at least the total number of firms in the sample times the applied significance level of the LM or Binomial test.

Figure 2 Scatter Diagram of DGP (11)[a]



[a]Crosses indicate relatively efficient companies whereas triangles indicate the inefficient ones. Output and input are on the y- and x-axes respectively.

$E = \{251,..,500\}$ contains 250 companies. The inefficient companies $(i \in \{1,..,250\})$ partly catch up with the others from period 1 to period 5 since the the added negative inefficiency component increases from $(-5-u_i)$ to $(-1-u_i)$. Note that the data generated by (11) seems to suit with the SFA estimation technique.

Table 1 contains the estimation results of both SFA and RTFA.[15]

The performance of SFA is clearly unsatisfactory because Classical Maximum Likelihood regression returns an estimate (1.129) which is relatively far from the parameter $\beta = 1$ (t-value of 229.24). This parameter estimate is biased because half of the observations are generated by a process which does not include the half normal distribution and because the mean of the added negative inefficiency component changes over time. This example highlights the extreme sensitivity of SFA to the distributional assumptions made by the researcher. Although the inefficiency terms are indeed generated by the half-normal distribution, model (1), even when it is specified with a half-normal component u_i and a normal error term v_i, does not exactly describe the data generating process (11). Evidently, in practice one cannot know the generating mechanism and this, together with a lack of powerful testing

[15] Our SFA algorithm uses the Newton-Raphson procedure when maximizing the likelihood function. These results can be reproduced by choosing the seed of the GAUSS random number generator equal to one.

Table 1 SFA and RTFA Estimation Results[a]

	SFA	RTFA (Stop when $\lambda^*_{LM} < \chi^2_{0.01}(10)$) Breusch-Pagan Test	RTFA (Stop when $\lambda^*_B < \chi^2_{0.01}(1)$) Binomial Test
β^*	1.129	0.985	0.979
	(229.24)	(334.05)	(322.08)
λ^*_{LM}		21.24	27.79
λ^*_B		4.84	3.27
No. of firms on the frontier		280	287
		Average X-efficiency	
Period 1		0.721	0.725
Period 2		0.770	0.774
Period 3		0.821	0.826
Period 4		0.869	0.874
Period 5		0.919	0.925

[a]t-values are in parentheses, $\chi^2_{0.01}(1) = 6.63$, $\chi^2_{0.01}(10) = 23.21$.

procedures, makes SFA questionable in real data applications. In our simulation experiment, the average inefficiency of the firms will be over-estimated with 13 per cent due to a dramatic decay of SFA.

RTFA however produces a reliable parameter estimate close to one. Note that the concomitant standard error is smaller than in the case of SFA. However, a Monte Carlo experiment could provide a decisive answer to the question whether RTFA is more efficient than SFA and whether it outperforms SFA when measuring against the Mean Squared Error. The Lagrange Multiplier test and the "Binomial test" statistic correctly reject that all observations in the sample follow a similar pattern. It takes up to 38 (40) iterations until RTFA identifies a group of relatively efficient firms which contains 287 (280) members in the case that the "Binomial test" (LM test) is applied. The latter number is reasonably close to the number of elements in E according to DGP (11). That is, DGP (11) is generated with 250 relatively efficient companies. Note that the moving averages of the inefficiency component in DGP (11) do not obscure RTFA. Indeed, the average X-efficiency is correctly estimated since the computed average value of measure (10a) climbs from about 0.7 in period 1 to 0.9 in period 5.

7. Conclusion and agenda for future research

We conclude that our Recursive Thick Frontier Approach provides a reliable alternative to the classical Stochastic Frontier Approach to estimating production or cost functions. RTFA is less dependent on *a priori* distributional assumptions regarding the inefficiency component of the disturbances than SFA and thus is to be preferred to SFA in applied studies where it is usually difficult to make a distinction between randomness and inefficiency. Furthermore RTFA provides powerful testing procedures for the underlying assumptions of the model.

The RTFA procedure, as outlined in this paper, remains an *ad hoc* solution to estimating inefficiency as we did not elaborate on the questions whether RTFA may not converge or converge in wrong directions. It is for future research to formulate precise conditions under which RTFA can successfully be applied.

References

Aigner, D., Lovell, C.A.K., and Schmidt, P., 1977, "Formulation and Estimation of stochastic Frontier Production Function Models", *Journal of Econometrics*, 6, 21-37.

Aigner, D.J., and Chu, S.F., 1968, "On the estimating the industry production function", *American Economic Review*, 58, 826-839.

Berger, A.N., 1993, " "Distribution-Free" Estimates of Efficiency in the U.S. Banking Industry and Tests of the Standard Distributional Assumptions", *The Journal of Productivity Analysis,* 4, 261-292.

Berger, A.N. and Humphrey, D.B., 1992, "Measurement and Efficiency Issues in Commercial Banking", in *Measurement Issues in the Service Sector,* Z. Griliches (ed.), NBER, Chicago.

Breusch, T.S., and Pagan, A.R. , 1980, "The Lagrange Multiplier Test and Its Applications to Model Specification in Econometrics", *Review of Economic Studies*, 47, 239-253.

Charnes, A., Cooper, W., Lewin, A.Y. and Seiford, L.M., 1994, *Data Envelopment Analysis*, Kluwer, Dordrecht, The Netherlands.

Charnes, A., Cooper, W.W., Rhodes, E., 1978, "Measuring the Efficiency of Decision Making Units", *European Journal of Operational Research*, 2, 6, 429-444.

Donoho, D.L. and Huber, P.J., 1983, "The notion of breakdown point", in *A Festschrift for Erich Lehmann*, edited by P. Bickel, K. Doksum and J.L. Hodges, Jr., Wadsworth, Belmont, CA.

Hinloopen, J. and Wagenvoort, J.L.M., 1997, "On the Computation and Efficiency of a HBP-GM Estimator: Some Simulation Results", *Computational Statistics and Data Analysis*, vol. 25, no. 1, 1-15.

Jondrow, J., Lovell, C.A.K., Materov, I.S. and Schmidt, P., 1982, "On the Estimation of Technical Inefficiency in the Stochastic Frontier Production Function Model", *Journal of Econometrics,* 19, 233-238.

Koenker, R.W. and Bassett, G.W., 1978, "Regression Quantiles", *Econometrica*, 46, 33-50.

Kopp, R.J. and Mullahy, J., 1990, "Moment-Based Estimation of Stochastic Frontier Models", *Journal of Econometrics*, 46, 165-183.

Rousseeuw, P.J. and Leroy, A.M., 1987, "Robust Regression and Outlier Detection", *Wiley*, New York.

Schure, P. and Wagenvoort, J.L.M., 1999, "Economies of Scale and Efficiency in European Banking: New Evidence", Economic and Financial Reports, 99/01, EIB.

Annex An Algorithm for the Recursive Thick Frontier Approach

RTFA

Initialisation:

Step 1: Set $j = 0$ and choose the speed of the data reduction process (for instance $\delta = 0.01$, i.e. in step 4 the data set is reduced with $\delta T * 100\%$).

Iteration:

Step 2: *Robust Estimation*
Compute the TLS estimates for $(1 - j * \delta * T) * 100\%$ of the data.

Step 3: *Test on Autocorrelation (or compute the Binomial test statistic λ_B)*
Compute Breusch and Pagan (1980) Lagrange Multiplier test statistic for the regression residuals associated with the TLS regression of step 2. If $\lambda_{LM} < CHI_{0.01}(T(T-1)/2)$ (99^{th} percentile of the chi-squared distribution with $(T(T-1)/2)$ degrees of freedom) then stop the iterations and report the last TLS regression results. Otherwise go to step 4.

Step 4: *Selecting the Relative Efficient Firms*
Compute the mean of the residuals $r_{ti} = y_{ti} - x_{ti}\beta^*$ for *each* cross-sectional unit i, $(m_i = mean(r_{1i},..,r_{Ti}))$. Sort the data on m_i. Set $j = j + 1$. Discard $j * \delta * T * 100\%$ of the observations by selecting $j * \delta * 100\%$ of the cross-sectional units with the smallest (largest) m_i in case of the production (cost) function. Repeat steps 2-4.