University of Victoria
Department of Economics

Econometrics Working Paper EWP0503

ISSN 1485-6441

A RECURSIVE THICK FRONTIER APPROACH TO ESTIMATING PRODUCTION EFFICIENCY

Rien Wagenvoort
&
Paul Schure

European Investment Bank, Luxembourg
&
University of Victoria, Canada

March, 2005

Abstract

We introduce a new panel data estimation technique for cost and production functions: the Recursive Thick Frontier Approach (RTFA). RTFA has two advantages over existing thick frontier methods. First, technical inefficiency is allowed to be dependent on the explanatory variables of the frontier model. Secondly, no distributional assumptions are imposed on the inefficiency component of the error term. We show by means of simulation experiments that RTFA can outperform the popular stochastic frontier approach (SFA) and the "within" OLS estimator for realistic parameterisations of the productivity model.

Keywords: Technical Efficiency, Efficiency Measurement, Frontier Production Functions, Recursive Thick Frontier Approach

JEL Classifications: C15, C23, C50, D2

* This paper was, in part, written while the second author visited the European Investment Bank in Luxembourg. The views expressed in this article are those of the individual authors and do not necessarily reflect the position of the EIB. We thank Jonathan Temple (the editor), two referees, Søren Johansen, Dermot O'Brien and seminar participants of the 1999 Econometric Society European Meeting in Santiago de Compostela for useful comments. Evidently, all remaining errors are ours.

Author Contact:
Paul Schure, Dept. of Economics, University of Victoria, P.O. Box 1700, STN CSC, Victoria, B.C., Canada V8W 2Y2; e-mail: schure@uvic.ca; FAX: (250) 721-6214

1. Introduction

The technical or X-efficiency of a firm measures the extent to which the firm's realised output is in line with the production frontier, i.e. the output realised by the "best-practice" firms as a function of the input bundle.[1] In the last 40 years a vast literature has attempted to tackle the difficult problem of establishing the production frontier. This paper introduces the Recursive Thick Frontier Approach (RTFA), a new panel data estimation method for estimating the production frontier. RTFA does not require a distributional assumption on the inefficiency component of the error term, and it allows technical inefficiency to be dependent on the explanatory variables of the frontier model. Unlike some of the other panel data methods, RTFA works well even if the number of time-periods in the panel dataset is small.

Traditional frontier estimation techniques can be divided into two groups, *full frontier* models and *thick frontier* models. Full frontier models assume that all deviations from the frontier represent inefficiency. Thus, all observations are lying on one side of the frontier. Schmidt (1975) shows that maximum likelihood estimation of full frontier models boils down to the linear or quadratic programming techniques introduced by Aigner and Chu (1968) for certain distributional assumptions on the inefficiency term. Greene (1980) discusses maximum likelihood estimation of full frontier models more generally. Nowadays, Data Envelopment Analyses (DEA) is the common name for the mathematical programming approach (Charnes, Cooper and Rhodes (1978) and Charnes, Cooper, Lewin and Seiford (1994)).

Thick frontier models or econometric frontier models assume that production levels may deviate from the frontier due to measurement errors or to factors beyond the control of the firm's management, besides inefficiency. Thus, observations may lie on both sides of the

frontier. Estimation procedures for thick frontier models have been developed for both cross-section and panel data. Popular cross-section methods are the Stochastic Frontier Approach (SFA) of Aigner, Lovell and Schmidt (1977) and Meeusen and Van den Broeck (1977), the Generalized Method of Moments of Kopp and Mullahy (1990), the Thick Frontier Approach (TFA) of Berger and Humphrey (1992), and the stochastic coefficients approach of Kalirajan and Obwona (1994). The probabilistic frontier production function of Timmer (1971) is a specific case of the stochastic coefficients approach. Schmidt and Sickles (1984), Battese and Coelli (1988), Cornwell, Schmidt, and Sickles (1990), and Kumbhakar (1990) introduce panel data methods to estimate the thick frontier.

RTFA is a thick frontier approach for panel data. RTFA hinges on the logical implication that if deviations from the frontier for the best-practice firms are random and symmetric around zero, then in a given time-period a best-practice firm is located either above or below the frontier with probability a half, independently of its location in other time-periods. This hypothesis can be tested in the case of panel data, but it requires sorting the sample into a subset of efficient, and a subset of inefficient firms. RTFA uses a recursive method to sort the data.

In a nutshell, RTFA starts with an Ordinary Least Squares (OLS) regression on the full sample of pooled observations. Each recursion proceeds with the computation of a Chow test statistic to test whether all firms can be considered as efficient. When the Chow test is rejected in iteration *j*, then the sample is reduced by eliminating all time-observations of the $\delta * j$ percent of the firms with the lowest residuals. The next iteration repeats the regression and the Chow test on the reduced sample. Observe that in each RTFA iteration the sample size is reduced by δ percent of the firms. Each time the regression and the Chow test are

[1] The cost frontier represents the cost level of the best-practice firms in the sample as a function of the output bundle and input prices. The exposition in this paper focuses on production frontiers, but all that follows applies

based on the reduced sample, while the full sample is considered when selecting the observations to be discarded. Thus, observations that were discarded in earlier iterations are reconsidered. Reconsidering previously discarded observations is important because the estimate of the "frontier" may change with each iteration. The RTFA algorithm stops once the Chow test fails to reject the hypothesis that all firms in the reduced sample are efficient. The RTFA frontier parameter estimates are the OLS estimates of the final iteration.

The key assumptions behind RTFA are weak in comparison to the assumptions of traditional techniques. The core assumption is that there exists a group of firms that are technically efficient in each time-period of the data. This seems a reasonable assumption when the dataset covers a "short" time span. In a short period of time it is unlikely that management undergoes a major change or that new technologies get implemented. It is therefore reasonable to assume that differences over time in measured productivity are due to "random errors" (measurement errors, good luck or bad luck, market conditions, etc). However, when the dataset covers a longer time span, this assumption may be strong. Over a longer time-period new technologies may be developed and implemented by a new management team. If there are no (or hardly any) firms in the dataset that are efficient in each time-period then RTFA cannot be applied directly. In this case we recommend splitting up the sample into shorter sub-periods and applying RTFA to each sub-period separately.

RTFA has two appealing features that follow directly from the result that only firms classified as efficient determine the frontier. First, RTFA permits correlation between inefficiency levels and inputs. Secondly, the frontier parameters are allowed to be different for efficient firms and inefficient firms. This is a desirable property because "the frontier function ... may *not* be a neutral transformation of the average function" [Timmer (1971), page 779].

---

equally well to cost frontiers.

RTFA can perhaps best be compared with the Thick Frontier Approach (TFA) of Berger and Humphrey (1992). TFA applies OLS to the quarter of the observations with the highest average production. In contrast to TFA, RTFA does not impose an assumption that the dataset has predetermined proportions of efficient and inefficient firms. In addition, RTFA sorts firms on the basis of their distance to the regression line, while TFA looks at average production. As a consequence, in the case of increasing returns to scale, TFA tends to omit small efficient firms, while large efficient firms tend to be omitted in the case of decreasing returns to scale.

RTFA does not share the two main drawbacks of the SFA production model. First, SFA requires *a priori* distributional assumptions regarding the error terms, which are difficult to test, and which affect the estimated frontier and therefore also the inefficiency estimates.[2] RTFA does not require a specification of the inefficiency term. Secondly, SFA hinges on the assumption that the inefficiency term is independent of the explanatory variables. For example, a significant relationship between firm size and inefficiency will bias the SFA estimates, unless the model is scaled by a size variable. The same applies for data exhibiting significant correlation between product mix and production efficiency. This problem is of course harder to fix than the effect of economies of scale, since it is *ex ante* difficult to predict how the output mix affects efficiency. The independency assumption of SFA is a serious weakness because it is violated for many real datasets. As observed by Førsund (1985-86, p. 329): "On account of these empirically observed (efficiency) differences it may then be unfortunate to assume efficiency differences neutral of the basic relationship between

[2] A sizable literature discussing the error term specification of SFA has developed. Stevenson (1980) and Greene (1990) assume that the inefficiency terms are distributed truncated normal and Gamma, respectively. Van den Broeck, Koop, Osiewalsi, and Steel (1994) adopt the Baysian approach, so that merely weak assumptions on the inefficiency term are imposed. Kopp and Mullahy (1990)'s Generalized Method of Moments estimation procedure enables various degrees of distributional flexibility and provides moment-based specification tests. They specify a parametric relationship between the first and third moments of the inefficiency term. This specification is *ad hoc* in itself, but their procedure enables testing of the validity of the distribution of the one-sided error component.

inputs and output." RTFA is not vulnerable to the criticism on SFA just explained. RTFA allows technical inefficiency to be dependent on the explanatory variables of the frontier model.[3]

Standard panel data approaches, such as the fixed effects and the random effects models, and RTFA share the advantage that no specific distributional assumptions have to be made regarding the distribution of the inefficiency term.[4] However, in the fixed effects model this comes at the cost of the assumption that inefficiency is time-invariant, and the problem that the estimate of the inefficiency term picks up both inefficiency and all other time-invariant firm-specific factors. In the case of the random effects model the inefficiency term and the regressors are assumed independent (unless the model is estimated with instrumental variables). Cornwell et al. (1990) do allow inefficiency to change over time. Their approach however imposes a functional form for the time pattern in inefficiency levels and only works when many time observations are available.[5] Battese and Coelli (1988) use ML estimation of the frontier, however this means that distributional assumptions on the error term have to be made.

Finally, RTFA and the Stochastic Varying Coefficients Frontier Approach (SVFA) of Kalirajan and Obwona (1994) have in common that the production model may be different for efficient and inefficient firms. A drawback of SVFA is that either panel data with many time observations must be available, or that many ad hoc parameter restrictions need to be

[3] In the case of panel data there is an additional problem with SFA. SFA estimation on the pooled data may bear out that the regression residuals are approximately normally distributed (Schure, Wagenvoort and O'Brien (2004) find this feature in panel data of European banks). In this case, the researcher would be tempted to conclude that the one-sided error component is negligible and therefore that the companies in the sample are equally efficient. However, one may find at the same time, that the residuals of individual companies (too) often have the same sign in each year of the sample period. In this case many companies have either persistently higher or persistently lower production than others. In other words, there are differences in production efficiency while this is not revealed by SFA.

[4] The fixed effects model and "within" OLS estimator are sometimes referred to as the Distribution Free Approach (DFA) in the context of efficiency studies (see Berger, 1993).

[5] Cornwell et al. (1990) assume that firm inefficiency is quadratic in time.

made. RTFA simply makes no assumptions on the technology used by inefficient firms. RTFA works with panel data that can be as short as two periods.

RTFA is explained in detail in the next section. In section 3 we contrast the performance of RTFA with standard OLS, SFA, and the "within" OLS estimator[6] in several simulation experiments. Section 4 concludes.

2. The Recursive Thick Frontier Approach to Estimating Technical Efficiency

Suppose there are n cross-sectional units ("firms") indexed by $i = 1,..,n$, and T time-periods indexed by $t = 1,..,T$, so that the full sample contains nT observations. Let the set of firms $N = \{1,...,n\}$ be comprised of two subsets E and H, the sets of technically efficient and technically inefficient firms, respectively. Consider the linear panel data model

$$y_{ti} = c_i + \alpha + x_{ti}\beta + \varepsilon_{ti}, \text{ where } c_i = 0, \quad i \in E. \tag{1}$$

This model describes the relationship between output y_{ti} and a k-dimensional input bundle x_{ti} for *technically efficient* firms only. As usual, α is an unknown constant, β is a k-dimensional column vector of unknown parameters and ε_{ti} is the error term of firm i in period t. The error term is random and does not reflect technical inefficiency. For inefficient firms the relationship between output and inputs remains unknown but, on average, inefficient firms are located below the production frontier:

$$\frac{1}{T}\sum_{t=1}^{T}(y_{ti} - x_{ti}\beta) = c_i < 0, \quad i \in H. \tag{2}$$

[6] Within estimates are computed by, first, subtracting the individual means from the observations (i.e. $\frac{1}{T}\sum_{t=1}^{T} y_{ti}$ from the observations on the dependent variable and $\frac{1}{T}\sum_{t=1}^{T} x_{ti}$ from the observations on the independent variables) and, then, applying OLS on the transformed data. The fixed effects are eliminated by the transformation of the data but can be easily estimated once the estimates of the parameters on the time-varying explanatory factors are obtained. See, among others, Baltagi (2001).

Notice that the degree of inefficiency of inefficient firms can be time variant.

We make the following assumptions:

Assumptions:

(A.1) For $i \in E$ and $t = 1,...,T$, ε_{ti} are independently and identically distributed with a distribution that is symmetric around zero.

(A.2) For $i \in E$ and $t = 1,...,T$, the orthogonality condition $E[\varepsilon_{ti} x_{ti}] = 0$ holds.

Assumptions (A.1) and (A.2) are standard assumptions of OLS. Notice, however, that we only make these assumptions for efficient firms. For the full sample of observations the assumptions may not hold, so that OLS should not be used.

Assumption (A.1) implies that the probability that an efficient firm is located above or below the frontier in a given time-period t is equal to one half, independently of its location in any preceding periods. In a panel data framework one can use a Chow statistic λ_{Chow}, to test whether the fixed effects c_i are zero for all firms in the dataset (see e.g. Baltagi, 2001). More precisely, we test whether $n-1$ fixed effects are equal to zero since model (1) includes a constant. The Chow test is based on the OLS residuals of the pooled model (i.e. without fixed effects) and the residuals of a "within" regression (i.e. including the fixed effects). It is a standard result that the asymptotic distribution of λ_{Chow} is a F-distribution with $(n-1, n(T-1)-k)$ degrees of freedom under the null hypothesis of no fixed effects.

RTFA begins with an OLS regression using the full sample of observations. If the Chow statistic is larger than the $(1-\theta)$th percentile of the F-distribution then we reject the hypothesis that all firms in the sample are efficient and we reduce the sample. In practice, we eliminate all time observations of $\delta\%$ of the firms with the lowest time-average of the residuals. We next repeat the regression and computation of the Chow test statistic on the

reduced sample. We continue to reduce the sample until the Chow test fails to reject assumption (A.1), i.e. until the largest possible group of efficient firms has been identified. The RTFA estimator is the OLS estimator of the last iteration of the algorithm. The details of the RTFA algorithm are given below.

The RTFA algorithm:

Initialisation:

Step 1: Set $j = 0$. Let the *current sample* be the full sample. Choose δ, i.e. the speed of the data reduction process, and θ, the significance level applied to the stopping criterion. For instance, $\delta = 0.01$ means that 1 percent of the firms are discarded in each iteration.

Iteration [The iteration starts with a current sample of $(1 - j * \delta) * 100\%$ of the data.]:

Step 2: *OLS Estimation*

Compute $\hat{\beta}_{j,OLS}$, the OLS estimator of the pooled model (the model without fixed effects)

Step 3: *Test Assumption (A.1)*

Compute the "within" estimates of the model with fixed effects. Use the restricted residual sums of squares (RRSS) of the OLS regression of step 2 and the unrestricted residual sum of squares (URSS) of the "within" regression to construct the test statistic λ_{Chow}. If λ_{Chow} is smaller than the $(1-\theta)$ th percentile of the F-distribution, then go to Step 5. Otherwise go to Step 4.

Step 4: *Select the Relatively Efficient Firms*

Compute the mean $(m_i = mean(r_{1i},..,r_{Ti}))$ of the residuals $r_{ti} = y_{ti} - x_{ti}\hat{\beta}_{j,OLS}$ for *each* cross-sectional unit i, including the firms that were discarded in previous iterations. Sort the data on m_i. Set $j = j + 1$. Discard $j * \delta * 100\%$ of the observations by selecting $j * \delta * 100\%$ of the cross-sectional units i with the smallest m_i. Go to step 2.

End the algorithm:

Step 5: The RTFA estimator

Define the RTFA estimator: $\hat{\beta}_{RTFA} = \hat{\beta}_{j,OLS}$. That is the RTFA estimator is the OLS estimator computed in Step 2 of the last iteration.

Notice that the final RTFA estimates are the OLS estimates of a model without (firm-specific) fixed effects. In the RTFA algorithm the "within" estimator is only used to compute the Chow test statistic. In the case of the classical "within" estimator inefficient firms may influence the slope parameters of equation (1), while this is not the case for RTFA.

The significance level θ applied to λ_{Chow} must be chosen in such a way that a distinction between randomness and inefficiency becomes relevant. In general, too low (high) choices of θ tend to leave (omit) too many inefficient (efficient) firms in the sample and RTFA tends to return biased estimates. There is no obvious optimal rule to follow when choosing θ because there is no such thing as an optimal significance level for a statistical test. We suggest to routinely carry out a sensitivity analysis based on alternative choices for θ when using RTFA.

The speed of the data reduction process δ is best chosen when to be as low as possible. In fact, since the RTFA algorithm runs very fast, we suggest setting δ equal to $1/n$. In this case the number of firms that are discarded increases one-to-one with the number of iterations.

Once the frontier is established firm i's average performance over time is computed as[7]:

$$XEFF_i = \frac{1}{T}\sum_{t=1}^{T}(y_{ti} / x_{ti}\,\hat{\beta}_{RTFA}), \text{ for } i \notin E$$

$$XEFF_i = 1, \text{ for } i \in E. \qquad (3)$$

[7] One may also compute the average performance of inefficient firms ($i \notin E$) in period t. However, the efficiency of cross-sectional unit i ($i \notin E$) in period t cannot be estimated with RTFA unless assumptions are made about the distribution of $XEFF_{ti}$ such as the SFA model assumptions.

The efficiency measure $XEFF_i$ indicates whether the firm optimally uses its resources x_{ti}. In other words, equation (3) measures X-efficiency, i.e. the firm's technical efficiency or managerial efficiency, rather than scale or scope efficiency.

Model (1) considers a production function with a *single output*. When the firm produces multiple outputs RTFA can be used to establish the *cost frontier* instead of the production frontier. Under weak conditions on the best-practice firms the principle of duality applies so that the cost function represents the production technology just as well as the production function.

The core assumption of RTFA is that there exists a group of firms that are technically efficient in each time-period of the data. This assumption may be strong when the dataset spans many time-periods. In this case we recommend splitting up the sample in subsets of at least two time-periods and applying RTFA to each subset. It is difficult to decide *a priori* how to break up long time-series and we suggest again that the researcher carries out a sensitivity analysis.

Finally, two pitfalls that are relevant for all thick frontier estimators have to be taken into account for RTFA as well. First, serial correlation in the errors of the efficient firms is a warning that the production model is not well specified. It generally implies that the frontier is shifting over time (e.g. because of technological progress). This problem may be resolved by introducing time dummies that pick up structural factors over time. The second pitfall concerns the adjustment for outlying observations. While outliers are problematic in general, they are even more alarming for frontier models (see e.g. Timmer, 1971). RTFA eliminates the inefficient firms step by step, including outliers that lie below the production frontier. However, outliers that are positioned *above* the production frontier may still push the frontier too far up. In order to obtain outlier-robust estimates we recommend the use of robust

estimators such as a High Breakdown Point Generalised M (HBP GM) technique instead of OLS (see e.g. Simpson et al. (1992) or Hinloopen and Wagenvoort (1997)). Simar (2003) proposes to carry out, in a first step, an exploratory data analysis to detect "super-efficient" outliers before using any frontier estimation method. The outlier detection tool introduced by Simar (2003) uses the nonparametric "order-*m* frontier" estimator of Cazals et al. (2002).

3. Simulation Study Results

We evaluate the relative performance of the Recursive Thick Frontier Approach with respect to OLS, the "within" estimator, and the traditional Stochastic Frontier Approach (SFA) through Monte Carlo simulations. OLS serves as a benchmark because it is the optimal approach when all firms in the sample are efficient. When there are differences in efficiency then OLS is biased. In that case, the within estimator is consistent (though not necessarily efficient) provided that the inefficiency component is independent of the explanatory variables. We compare RTFA with SFA since SFA is the most widely applied technique among the thick frontier approaches. Our setup consists of a labour productivity model where productivity is log-linear in the capital-to-labour ratio and inefficiency.

Below we will first assess the performance of the above-mentioned estimators in estimating the frontier parameters using Mean Squared Error (MSE) as our main criterion. After that we compare the distribution of the X-efficiency estimates and the distribution of the true X-efficiencies.

3.1. Design of the experiments

We generate labour productivity data of 500 firms for five time-periods. The data generating process is the following:

$$y_{ti} = x_{ti}^{\beta} EXP(\varepsilon_{ti}), \quad t = 1,..,5, \quad i = 1,..,500 \qquad (4)$$

such that

$$x_{ti} = 10 + 10\eta_{ti}, \tag{5}$$

$$\varepsilon_{ti} = v_{ti} + u_{ti}. \tag{6}$$

Here y_{ti} represents the output of firm i per unit of labour in period t and x_{ti} is the capital-to-labour ratio. Thus, apart from disturbances we have a very simple production function for which labour productivity increases by β percent when the capital-to-labour ratio increases by 1 percent. The error term ε_{ti} consists of an inefficiency component, u_{ti}, and a random component, v_{ti}.

Throughout the simulation study we assume that η_{ti} is a random draw from a standard half-normal distribution; that v_{ti} is normally distributed with zero mean and variance equal to 1/9; and that $\beta = 1$ for all firms i. Regarding the inefficient component of the disturbances, we study three different cases. In Case 1 we follow the SFA assumptions of Aigner et al. (1977). In Case 2 we examine the impact of a small perturbation in the SFA assumptions and set the inefficiency component u_{ti} at zero for efficient firms. In Case 3 we assume that inefficiency is related to the capital-to-labour ratio. This clearly violates the SFA assumptions. In all three cases we estimate the model in logs.

Case 1: The SFA assumptions

The data generating process of (4)-(6) becomes the SFA model of Aigner et al. (1977) when taking logs of both sides of equation (4); assuming that *each* individual firm suffers from the inefficiency term u_{ti} (i.e. $H = N$ and $E = \varnothing.$); and u_{ti} is generated by a standard (negative) half normal distribution,

$$u_{ti} \leq 0, \; i \in N.$$

In Case 1, firms meet the output targets set on the basis of the production function with probability zero. Indeed, each firm i is, on average, $\frac{1}{T}\sum_{t=1}^{T} u_{ti} \neq 0$ away from potential labour productivity. For many real-word datasets, one may *a priori* expect that at least some firms are actually efficient. Cases 2 and 3 provide examples of such more realistic scenarios.

Case 2: Modified SFA assumptions. The inefficiency term is zero for efficient firms.

In Case 2 we assume that the first 250 firms are fully efficient in all five periods. The inefficiency component of the 250 inefficient firms follows a standard (negative) half-normal distribution in each time-period. Thus,

$$u_{ti} = 0,\ i \in E = \{1,..,250\}$$

$$u_{ti} \leq 0,\ i \in H = \{251,..,500\}.$$

Case 3: Inefficiency is negatively related to capital intensity

In Case 3 there is a negative relationship between inefficiency and the capital-to-labour ratio. In other words, inefficient firms that are capital-intensive are relatively more efficient than labour-intensive firms. In particular,

$$u_{ti} = 0,\ i \in E = \{1,..,250\}$$

$$u_{ti} = \ln\left(\frac{0.5x_{ti}}{\bar{x}}\right),\ i \in H = \{251,..,500\},$$

where $\bar{x}$ is the average of the time-averaged capital-to-labour ratio of all inefficient firms. For instance, an inefficient firm i with an average capital-to-labour ratio in period t (i.e. $x_{ti} = \bar{x}$) is 50% inefficient since its labour productivity is half the corresponding ratio of an efficient firm with an equal capital intensity. Similarly, an inefficient firm with half the average capital-to-labour ratio in period t (i.e. $x_{ti} = 0.5\bar{x}$) is 75% inefficient. A firm with twice or more the average capital-to-labour ratio in period t (i.e. $x_{ti} \geq 2\bar{x}$) is *not* inefficient.

Therefore, although the likelihood of this event is small, some of the firms in H could actually be positioned above the production frontier (on average).

We report statistics for each case based on ten thousand successfully completed trials. When implementing RTFA we chose the speed of the data reduction equal to 1 firm per iteration ($\delta = 1/n$) and applied a significance level of $\theta = 0.05$ to the Chow test statistic. To compute the SFA estimates we used the Newton-Raphson algorithm to find the maximum likelihood estimators for β and λ, i.e. the ratio of the standard deviation of u to the standard deviation of v. As starting conditions for the Newton-Raphson algorithm we chose $\beta = \hat{\beta}_{OLS}$ and $\lambda = 0.5$. Two convergence criteria are applied when running SFA. Firstly, the percentage change in the parameter estimates between the final and next to last iteration must be less than 0.01%. Secondly, the absolute value of each entry of the gradient must be smaller than 0.0001. We terminated the Newton-Raphson algorithm when we obtained complex numbers or the number of iterations exceeded 100.

3.2 Estimating the elasticity of labour productivity

Table 1 shows the simulation results regarding our estimations of β, the elasticity of labour productivity. The top, middle, and bottom panel of Table 1 show the results for Case 1, 2, and 3 respectively. The first row of each panel contains the mean of the parameter estimates for β over 10,000 successfully completed trials. The remaining rows of each panel show the median, the minimum and maximum value over the 10,000 runs, as well as the variance and the Mean Squared Error of the parameter estimates.

3.2.1 Case 1: The SFA assumptions

Table 1 shows that SFA outperforms RTFA according to the MSE criterion. This is an unsurprising result since SFA provides the minimum variance unbiased estimation technique under the SFA assumptions (see Aigner et al., 1977). RTFA breaks down in case 1 because it

is unable to distinguish between efficient and inefficient firms. Indeed, the rounded average of the number of firms that is selected as efficient by RTFA is equal to 500. Therefore, RTFA effectively becomes OLS for a typical trial, which explains why the RTFA results are very similar to the OLS findings. RTFA and OLS are clearly biased. RTFA fails because in each period t the set of efficient firms is completely different.

3.2.2 Case 2: Modified SFA assumptions. The inefficiency term is zero for efficient firms

In Case 2 the inefficiency term is zero for efficient firms throughout the sample period. Table 1 shows that RTFA outperforms SFA in terms of the MSE. RTFA, on average, comes very close to the true value of $\beta = 1.$ SFA overestimates β by 13% on average. The number of firms predicted to be efficient by RTFA is a bit higher than the actual number of best-practice firms. Note b of Table 1 shows that RTFA classifies on average 273, instead of 250, firms as efficient. The fact that RTFA removes too few firms on average is the reason why the elasticity parameter β is slightly under-estimated. In the first run of the simulation experiment we found that all truly efficient firms were indeed used by RTFA to estimate the frontier. We did not check this for the remaining 9,999 rounds. In Case 2 all RTFA estimates are found in the [0.97, 1.01] range, while SFA estimates vary between –0.97 and 5.47.

The literature has shown that SFA is extremely sensitive to the distributional assumptions regarding the inefficiency component. Case 2 provides clear evidence of this problem. By simply assuming that efficient firms have an inefficiency term of zero SFA becomes far from consistent.

In Case 2, RTFA is also better than OLS and the "within" estimator in terms of MSE. In addition, OLS gives estimates that are on average biased by 14%. The within estimator is consistent, but has a substantially higher variance in the parameter estimates and is thus less efficient than RTFA.

3.2.3 Case 3: Inefficiency is negatively related to capital intensity

In Case 3, the true elasticity parameter of $\beta = 1$ is not found by the within estimator, SFA or OLS. The reason is that inefficiency and the capital-to-labour ratio are correlated. This violates the standard orthogonality condition. RTFA removes inefficient firms from the sample and is thus able to find the true parameter $\beta = 1$ on average. RTFA selects 262 firms on average, which is not far from the true number of 250 efficient firms. RTFA also outperforms the within estimator, SFA and OLS in terms of MSE.

Table 1. Estimation of the elasticity of labour productivity; A comparison beween RTFA, SFA, OLS, and the "within" estimator for cases 1, 2, and 3. True parameter value is $\beta = 1.$ [a]

	RTFA[b]	SFA[c]	OLS	Within OLS
	Case 1: The SFA assumptions			
Mean	0.72250	0.99998	0.72238	1.00064
Median	0.72246	1.00017	0.72233	1.00077
Minimum	0.70632	0.73181	0.70632	0.81933
Maximum	0.74393	1.02782	0.74393	1.17265
Variance	0.00002	0.00007	0.00002	0.00240
Mean Squared Error	0.07703	0.00007	0.07710	0.00240
	Case 2: Modified SFA assumptions. The inefficiency term is zero for efficient firms			
Mean	0.99180	1.12972	0.86117	0.99977
Median	0.99180	1.13609	0.86117	1.00041
Minimum	0.97444	-0.96811	0.84558	0.85163
Maximum	1.00937	5.46939	0.87498	1.14533
Variance	0.00003	0.00504	0.00001	0.00143
Mean Squared Error	0.00009	0.02187	0.01929	0.00143
	Case 3: Inefficiency is negatively related to capital intensity			
Mean	0.99537	1.08581	0.87654	1.49986
Median	0.99537	1.08585	0.87654	1.49999
Minimum	0.97607	1.06056	0.86757	1.40255
Maximum	1.01409	1.10939	0.88612	1.60089
Variance	0.00002	0.00004	0.00001	0.00073
Mean Squared Error	0.00004	0.00740	0.01525	0.25059

[a] Statistics are based on ten thousand successfully completed iterations for each estimation procedure.
[b] The average number of "efficient firms" as determined by RTFA is equal to 500, 273, and 262 in cases 1, 2, and 3 respectively. RTFA did not fail in any of the first 10,000 iterations.
[c] SFA failed to converge 13,179, 13,585, and 0 times in cases 1, 2, and 3 respectively, so that the total number of iterations was 23,179, 23,585, and 10,000, respectively.

3.3 Estimating X-efficiency

Let us now contrast the distributions of the X-efficiency estimates of RTFA, SFA, OLS, the within estimator, and the true X-efficiencies. To do this we computed, for each run, the average X-efficiency of each firm over the five time-periods. We obtained the true average X-efficiency of firm i by ignoring the random component v_{ti} and setting u_{ti} to its true value:

$$XEFF_{i,TRUE} = \frac{1}{T}\sum_{t=1}^{T}\frac{x_{ti}EXP(u_{ti})}{x_{ti}} = \frac{1}{T}\sum_{t=1}^{T}EXP(u_{ti}) \qquad (7)$$

According to RTFA the average X-efficiency of firm i is given by:

$$XEFF_{i,RTFA} = \frac{1}{T}\sum_{t=1}^{T}(y_{ti} / x_{ti}^{\hat{\beta}_{RTFA}}) \text{, for } i \notin E$$

$$XEFF_{i,RTFA} = 1 \text{, for } i \in E \qquad (8)$$

In the case of SFA we use the result of Jondrow, Lovell, Materov, and Schmidt (1982) who show that the conditional distribution of u_{ti} given ε_{ti} is truncated normal when u_{ti} is half normal. In particular, we computed the expected value of u_{ti} as:

$$\hat{u}_{ti} = \frac{\sigma_u\sigma_v}{\sigma}\left[\frac{f(e_{ti}\sigma_u / \sigma_v\sigma)}{1-F(e_{ti}\sigma_u / \sigma_v\sigma)} - \frac{e_{ti}\sigma_u}{\sigma_v\sigma}\right] \qquad (9)$$

Here f and F represent the density function and the distribution function of the standard normal distribution, respectively; e_{ti} is the residual of firm i in period t; and σ, σ_v and σ_u are the estimated standard deviations of ε_{ti}, v_{ti} and u_{ti} respectively. The average X-efficiency is now computed as follows:

$$XEFF_{i,SFA} = \frac{1}{T}\sum_{t=1}^{T}EXP(\hat{u}_{ti}) \qquad (10)$$

By definition, X-efficiency is equal to 1 for all firms when applying OLS. Finally, in the case of the "within" OLS estimator, the average inefficiency of firm i is estimated as:

$$XEFF_{i,Within} = EXP(\hat{u}_i) \qquad (11)$$

where $\hat{u}_i = c_i - \max(c_i)$, and c_i is the estimated fixed effect corresponding to firm i.

The simulation results regarding the X-efficiency estimates are presented in Figure 1 and Table 2. Figure 1 shows the histogram of the X-efficiency scores over 10,000 successfully completed trials. That is, we divided the "X-efficiency domain" (i.e. [0%, 100%]) into 20 intervals (i.e. [0%, 2.5%), [2.5%, 7.5%), [7.5%, 12.5%), etc.) and estimated the relative frequency of finding a X-efficiency score in each bracket over the 10,000 trials. Table 2 contains the average of the distribution shown in Figure 1. Thus, Table 2 shows the average X-efficiency level over all periods, all firms and all trials.

3.3.1 Case 1: The SFA assumptions

Table 2 shows that, as expected, the SFA model produces an average X-efficiency score that is close to the true average X-efficiency level of 52%. We discussed above that RTFA fails in Case 1 because it does not manage to distinguish between efficient and inefficient firms. RTFA hence produces X-efficiency scores of 100% just like OLS. The "within" estimator estimates average X-efficiency at 47%. Although the "within" estimator is unbiased for β, it overestimates the maximum of the fixed effects associated with the most "efficient" firm due to the relatively short time-dimension of the generated data. As a consequence, X-efficiency is underestimated by 5% on average.

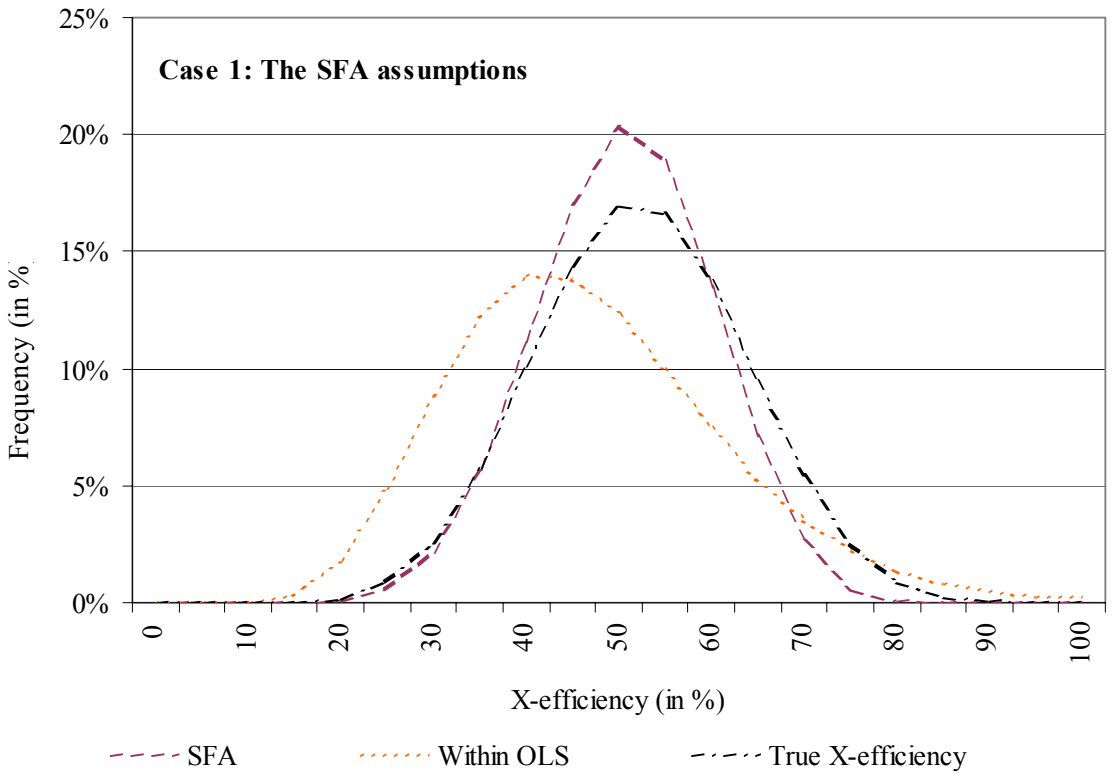
Figure 1 reveals that the distribution of the X-efficiency scores based on SFA is relatively close to the distribution of the true X-efficiency scores in Case 1. SFA tends to find slightly too few firms with a X-efficiency score between 60% and 100%, and slightly too many with

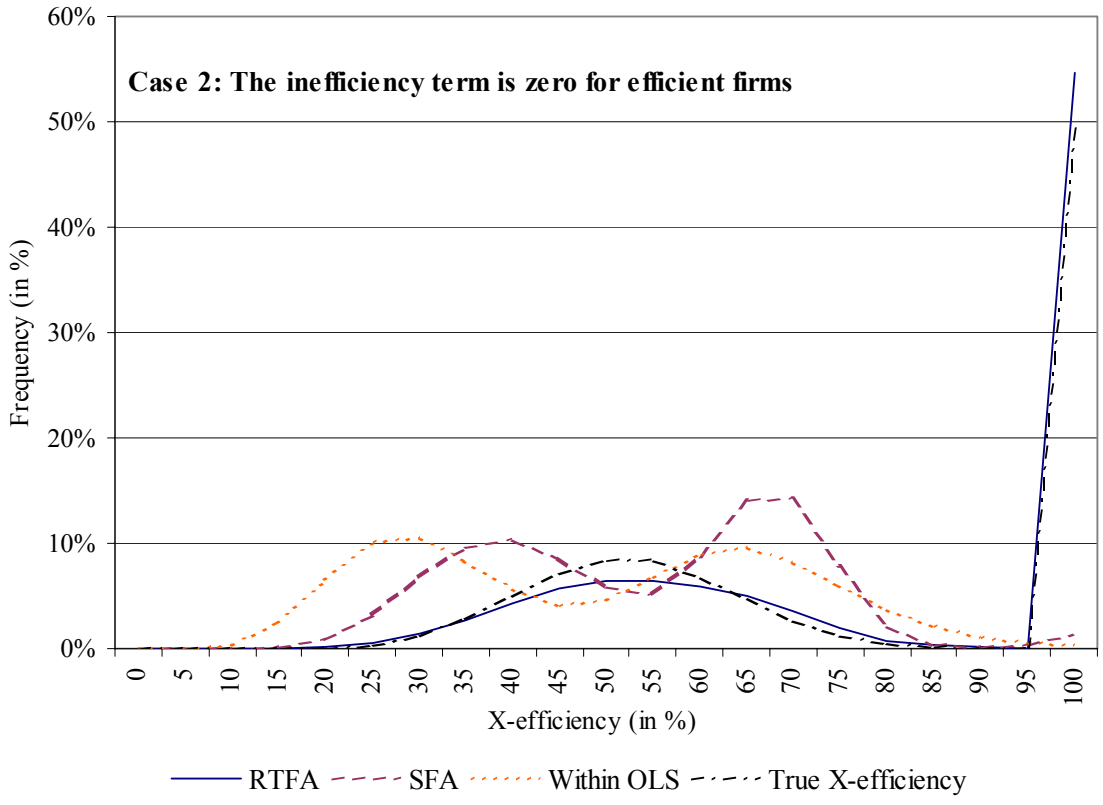
a X-efficiency score between 40% and 60%. The performance of SFA to estimate X-efficiency however contrasts favourably with the "within" estimator, which produces a distribution which is centred too far to the left.

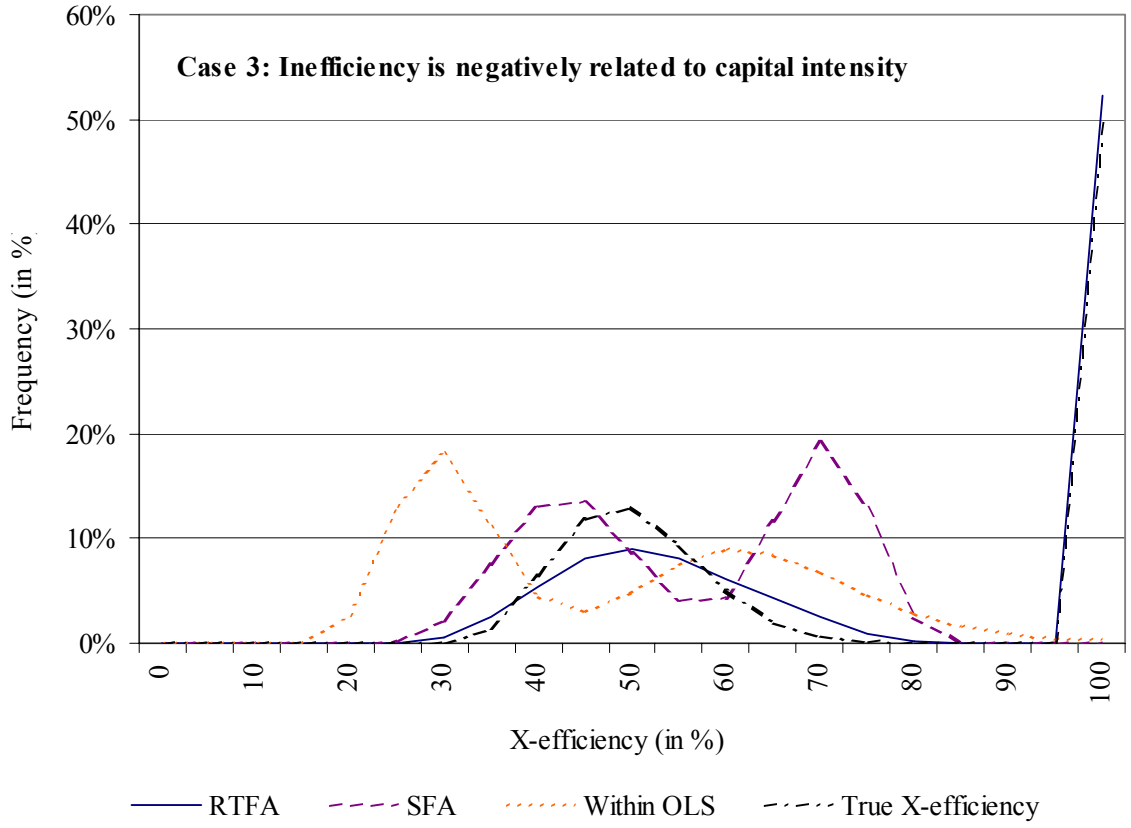
Table 2. Estimation of average X-efficiency; A comparison between RTFA, SFA, OLS, and the "within" estimator for cases 1, 2, and 3.

	RTFA	SFA	OLS	Within OLS	True
Case 1	100%	51%	100%	47%	52%
Case 2	79%	56%	100%	49%	76%
Case 3	77%	57%	100%	47%	75%

Figure 1. Distribution of average X-efficiency scores; A comparison between RTFA, SFA, and the "within" estimator for cases 1, 2, and 3.




Case 2: The inefficiency term is zero for efficient firms
60%
50%
40%
30%
20%
10%
0%
Frequency (in %)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100
X-efficiency (in %)
RTFA
SFA
Within OLS
True X-efficiency


Case 3: Inefficiency is negatively related to capital intensity
60%
50%
40%
30%
20%
10%
0%
Frequency (in %
0
10
20
30
40
50
60
70
80
90
100
X-efficiency (in %)
RTFA
SFA
Within OLS
True X-efficiency

3.3.2 Cases 2 and 3

The full sample average level of X-efficiency is considerably under-estimated by both SFA and the "within" estimator in the cases 2 and 3. For example, in Case 2, SFA and the "within" estimator find average scores of 56% and 49%, respectively, while the true score is 76% (Table 2). By contrast, RTFA produces an average X-efficiency score that is close, though slightly above, the true average X-efficiency score in cases 2 and 3.

In the previous subsection we showed that in cases 2 and 3 RTFA tends to classify slightly too many firms as efficient (Table 1). This result is also illustrated in Figure 1. In cases 2 and 3 the RTFA estimated frequency of fully efficient firms slightly exceeds the true frequency of 50%. However, overall the distribution of X-efficiency scores produced by RTFA is relatively close to the distribution of true X-efficiency scores. This cannot be said for SFA and the "within" estimator. They fail to find that actually 50% of the firms are X-efficient and, as a consequence, their histograms of the X-efficiency scores look entirely different altogether.

The simulation results clearly demonstrate that the bias in X-efficiency tends to be much larger than the bias in β. In Case 3, for example, SFA over-estimated β by about 8.6% on average (Table 1), while X-efficiency is underestimated by 24% on average (Table 2; note that (57-75)/75=-24%).

4. Conclusion

In this paper we introduced an intuitive method for the estimation of thick frontier models: the Recursive Thick Frontier Approach (RTFA). The key assumptions behind this new technique are weak when compared to the assumptions on which traditional econometric techniques are based. A strength of RTFA is that it uses only observations associated with "efficient firms" to estimate the frontier parameters. "Inefficient firms" do not influence the

estimates of the production function, and, unlike with traditional thick frontier techniques, inefficiency need not be parameterised in the production model.

We compare RTFA to OLS, the stochastic frontier approach (SFA) and the "within" panel data estimator through a number of Monte Carlo simulations. We find that, on the basis of the mean squared error of the parameter estimates over 10,000 trials, RTFA outperforms SFA, OLS and the within estimator for realistic parameterisations of the labour productivity model (i.e. cases 2 and 3). However, RTFA fails if in each period the set of efficient firms is different (Case 1). Our simulation results show that SFA is superior when its underlying distributional assumptions are met but SFA is not robust to small realistic departures from those assumptions.

References

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

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.

Baltagi, B. H. (2001). *Econometric Analysis of Panel Data*, Second edition, John Wiley & Sons, West Sussex, England.

Battese, George E. and Tim J. Coelli (1988). 'Prediction of Firm-level Technical Efficiencies with a Generalized Frontier Production Function and Panel Data', *Journal of Econometrics,* 38, 387-399.

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.

Broek, van den, Julien, Gary Koop, Jacek Osiewalski, and Mark F.J. Steel (1994). 'Stochastic Frontier Models: A Baysian Approach', *Journal of Econometrics,* 61, 273-303.

Cazals, C., Florens J. P. and Simar, L. (2002). 'Nonparametric Frontier Estimation: A Robust Approach', *Journal of Econometrics,* 106, 1-25.

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.

Cornwell, C., P. Schmidt, and R.C. Sickles (1990). 'Production Frontiers with Cross-Sectional and Time-series Variation in Efficiency Levels', *Journal of Econometrics*, 46, 185-200.

Førsund, F. R. (1985-86). 'Comment', *Econometric Reviews,* 4(2), 329-334.

Greene, W.H. (1980). 'Maximum Likelihood Estimation of Econometric Frontier Functions', *Journal of Econometrics*, 13, 27-56.

Greene, W.H. (1990). 'A Gamma-Distributed Stochastic Frontier Model', *Journal of Econometrics*, 46, 141-163.

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.

Kalirajan, K.P. and Obwona, M.B. (1994). 'Frontier production function: a stochastic coefficients approach', *Oxford Bulletin of Economics and Statistics*, 56, 85-94.

Kalirajan, K.P. and Shand, R.T. (1999). 'Frontier Production Functions and Technical Efficiency Measures', *Journal of Economic Surveys*, vol.13-2, 149-172.

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

Kumbhakar, S.C. (1990) 'Production Frontiers, Panel Data and Time-Varying technical Inefficiency', *Journal of Econometrics*, 46, 201-211.

Meeuwsen, W. and J. van den Broeck (1977). 'Efficiency estimation from Cobb-Douglas production functions with composed error,' *International Economic Review*, 18-2, 435-444.

Schmidt, P. (1975). 'On the Statistical Estimation of Parametric Frontier Production Functions', *Review of Economics and Statistics*, 58, 238-239.

Schmidt, P and R.C. Sickles (1984). 'Production Frontiers and Panel Data', *Journal of Business and Economic Statistics*, 2, 367-374.

Schure, P., J.L.M. Wagenvoort, and D. O'Brien (2004). 'The Efficiency and Conduct of European Banks: Developments after 1992', *Review of Financial Economics*, 13, 371-396.

Simar, L. (2003). 'Detecting Outliers in Frontier Models: A simple Approach', *Journal of Productivity Analysis,* 20, 391-424.

Simpson, D.G., Ruppert, D. and Carroll, R.J. (1992). 'On One-Step GM Estimates and Stability of Inferences in Linear Regression', *Journal of the American Statistical Association*, Vol. 87, no. 418, 439-50.

Stevenson, R.E. (1980). 'Likelihood Functions for Generalized Stochastic frontier estimation', *Journal of Econometrics*, 13, 57-66.

Timmer, C.P. (1971). 'Using a Probabilistic Frontier Production Function to Measure Technical Efficiency', *Journal of Political Economy,* 79, 767-794.