NHH

INSTITUTT FOR FORETAKSØKONOMI

DEPARTMENT OF BUSINESS AND MANAGEMENT SCIENCE

FOR 24 2015

ISSN: 1500-4066

September 2015

Discussion paper

On the Distributional Assumptions in the StoNED model

BY

Xiaomei Cheng, Jonas Andersson AND **Endre Bjørndal**

Norges Handelshøyskole

NORWEGIAN SCHOOL OF ECONOMICS

On the Distributional Assumptions in the StoNED model

Xiaomei Cheng, Jonas Andersson, Endre Bjørndal
Department of Business and Management Science
Norwegian School of Economics
Xiaomei.Cheng@nhh.no, Jonas.Andersson@nhh.no, Endre.Bjørndal@nhh.no

Abstract:

In a recent paper Johnson and Kuosmanen (2011) propose a new, semi-parametric, general cost-frontier model, the stochastic nonparametric envelopment of data (StoNED). The model is semi-parametric in the sense that the cost function is estimated non-parametrically, while the functional form of the distribution for the error term is parametrically specified. A common assumption for this distribution is that it is a convolution of a truncated normal distribution, representing inefficiency, and a normal distribution, representing noise. This parametric form has the drawback that a negative skewness implies a negative expected inefficiency. It can thus never capture a negatively skewed distribution with a positive expectation. In this paper we investigate this assumption and its consequences for an analysis of inefficiency. Furthermore, we propose a solution to the problem and investigate its performance by means of a Monte Carlo simulation.

Keywords: StoNED model; Composite error; Wrong skewness; Misspecification

1. Introduction

Until recently, there were two commonly used methods for studying efficiency performance of the electricity distribution sector: The deterministic, nonparametric Data Envelopment Analysis (DEA) (Charnes et al., 1978) and the parametric Stochastic Frontier Analysis (SFA) (Aigner et al., 1977). DEA is a non-parametric method in the sense that no functional form for the cost function needs to be specified. DEA is capable of handling both multiple inputs and multiple outputs. At the other end of the spectrum, SFA is fully parametric. A functional form for the cost function is specified together with an error term, consisting of the inefficiency and a noise term, with assumed probability distributions. The specification of the error term enables the modeler to investigate the model fit in a traditional econometric sense. Johnson and Kuosmanen (2011) proposed a method, which can be placed in between DEA and SFA on the non-parametric/parametric scale. It is called stochastic non-parametric envelopment of data (StoNED) and has the property that the cost function is estimated non-parametrically combined with a parametric assumption on the distribution of the error term, making it possible to separate inefficiency from noise. From 2012 this approach has been used for benchmarking and regulation of Finnish electricity distribution companies (Kuosmanen., 2012).

In parametric stochastic frontier models, a unit´s deviation from the efficient frontier is modeled as a sum of two components. These are the inefficiency component, which is modeled by a stochastic variable that only obtains positive values, and a random error that is typically modeled by a stochastic variable with a distribution that is symmetric around zero. In their seminal paper, Aigner et al. (1977) used the half-normal distribution for the inefficiency and the normal distribution for the random error. These random errors are in addition assumed to be independent and identically distributed across observations and statistically independent of each other. Other distributions that have been used for the inefficiency are the exponential, the truncated normal and two-parameter gamma distributions (Meeusen and van den Broeck, 1977; Aigner, Lovell and Schmidt, 1977; Stevenson, 1980; Greene, 1980). When adding these stochastic variables representing inefficiency and random error, an important feature can be observed. Both the expected value and the skewness are non-negative. This theoretical feature is however not always matched in the residuals obtained after fitting a stochastic frontier model. Green and Mayes (1991) showed that 48 of 151 industries in United Kingdom

had an unexpected sign of the skewness, and they also reported that 49 of 140 Australian industries had the same feature.

A common conclusion, when negative skewness is observed, is that the model is misspecified. While this is impossible to refute, a model should after all produce results that corresponds to its assumptions, it is not obvious which of the assumptions that are unfulfilled. In this paper we will argue, in line with Carree (2002), that the deviation from the efficient cost frontier might well have a negative skewness at the same time as a positive expectation. It is actually the latter, a positive expectation, that is essential for the economics of the model to make sense. A negative skewness is merely a technical inconvenience that makes some existing estimation procedures difficult to implement.

This weak point of stochastic frontier models has been studied in a number of papers. Simar and Wilson (2009) argued that "wrong skewness" was not an estimation or modeling failure, but a finite sample problem that is most likely to occur when the signal-to-noise ratio (the variance ratio of the inefficiency component to the variance of the composite error) was small. An interpretation of this is that the error term (inefficiency plus noise) is basically normally distributed. Theoretically, it therefore has a skewness of zero. In a sample, however, it can obviously happen to be somewhat negative. When the skewness is negative, two possible solutions are to obtain a new sample or to re-specify the model (Carree, 2002; Almanidis et al., 2011). Following Simar and Wilson (2009), Qu et al (2013) proposed a non-positivity residual skewness constraint in the maximum likelihood estimation (MLE) algorithm in order to avoid the problem. The disadvantage of this approach is that the ratio of signal to noise, which affects the results, could not be determined. The weakness can be avoided in panel data models if we use the fixed effects model proposed by Schmidt and Sickles (1984) or time-varying models developed by Cornwell et al (1990, 1996).

This paper will investigate an alternative approach not yet considered for the StoNED model. Our approach is based on the following observation: A negative skewness is only unreasonable if it implies inefficiencies with a negative expectation. Carree (2002) presented this idea for stochastic frontier models and exemplifies it with the binomial distribution. In the present paper, we will modify the two most commonly used distributions in the StoNED context, the half normal and the exponential distributions, so that they have a negative skewness and a positive expectation. We will also study how

robust the StoNED estimator of the cost function is with respect to distributional assumptions on the inefficiency.

The rest of this paper is divided into 5 sections. The next section discusses the StoNED model and Section 3 suggests a modification to it. Section 4 presents an illustration with electricity data and the results of a small Monte Carlo study investigating the sensitivity of the distributional assumptions for the modified StoNED model. Section 5 concludes.

2. The StoNED Model

In line with Kuosmanen (2012), we assume that the observed data is consistent with the relationship

$$x_i = C(\boldsymbol{y}_i) \cdot \exp(\varepsilon_i). \tag{1}$$

The observed total cost for a company is denoted x_i, C is the cost frontier function, and $\boldsymbol{y}_i$ is the vector of outputs of company i. The error term can be decomposed in two parts, $\varepsilon_i = u_i + v_i$, where v_i is a stochastic noise term and u_i represents inefficiency. The noise v_i is assumed to be normally distributed with a zero mean and a finite variance σ_v^2. The inefficiency u_i is usually assumed to follow a half-normal distribution $|N(0, \sigma_u^2)|$, with the variance $Var(u_i) = \frac{\pi-2}{\pi}\sigma_u^2$, the expected value of inefficiency $E(u_i) = \mu = \sigma_u\sqrt{2/\pi} > 0$ (Aigner et al., 1977). We also assume that v_i and u_i are statistically independent of each other and of the regressors.

As described by Kousmanen (2012), the StoNED method has two stages:
Stage 1: Estimate the total cost by convex nonparametric least squares (CNLS).

Stage 2: Estimate the variance parameters σ_u^2, σ_v^2, the expected values of inefficiency μ and the cost frontier function $\hat{C}$.

The CNLS estimator can be obtained by the following convex programming model:

$$\min_{\alpha,\beta,\varepsilon,\gamma} \sum_{i=1}^{n} \varepsilon_i^2$$

s.t.

$$\begin{aligned} &\ln x_i = \ln \gamma_i + \varepsilon_i && i = 1, \cdots, n \\ &\gamma_i = \alpha_i + \boldsymbol{y}_i\boldsymbol{\beta}_i' \geq \alpha_h + \boldsymbol{y}_i\boldsymbol{\beta}_h' && h = 1, \cdots, n \end{aligned} \tag{2}$$

$$\boldsymbol{\beta}_i \geq 0 \qquad\qquad i = 1, \cdots, n$$

Here, γ_i is the convex nonparametric least squares (CNLS) estimator of the expected total cost of producing the output vector $\boldsymbol{y}_i$, $\boldsymbol{\beta}_i$ is the vector of the marginal cost of outputs for firm i, and α_i is the intercept of firm i. The first constraint of model (2) can be interpreted as the regression equation. The second and third constraints ensure convexity and monotonicity, respectively. Model (2), where the sign of α_i is unrestricted, implicitly assumes variable returns to scale (VRS), and alternative scale assumptions can be expressed by imposing restrictions on the sign of α_i (Kuosmanen and Kortelainen, 2012).

For stage 2 of the StoNED procedure, there are two approaches to estimate the variance parameters based on the optimal solution $\hat{\varepsilon}_i$ of model (2): the method of moments (MoM) (Aigner et al., 1977) and the pseudo-likelihood estimation approach (PSL) (Fan and Weersink, 1996). We only consider the former method, since the computation is simpler than for the latter one. Then, under half-normal inefficiency and normal noise, the parameters of the two distributions can be obtained by

$$\hat{\sigma}_u = \sqrt[3]{\frac{\widehat{M}_3}{\left(\sqrt{\frac{2}{\pi}}\right)\left[\frac{4}{\pi}-1\right]}}, \text{and} \tag{3}$$

$$\hat{\sigma}_v = \sqrt[2]{\widehat{M}_2 - \left[\frac{\pi-2}{\pi}\right]\hat{\sigma}_u^2}, \tag{4}$$

where $\widehat{M}_2 = \sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\varepsilon})^2/n$ and $\widehat{M}_3 = \sum_{i=1}^{n}(\hat{\varepsilon}_i - \bar{\varepsilon})^3/n$ are estimates of the second and third central moments of the composite errors distribution, respectively. The cost frontier function is estimated by

$$\hat{C}(\boldsymbol{y}_i) = \gamma_i * \exp\left(-\hat{\sigma}_u\sqrt{\frac{2}{\pi}}\right). \tag{5}$$

Furthermore, the cost efficiency score of firm i is defined as

$$CE_i = \frac{\hat{C}(\boldsymbol{y}_i)}{x_i}, \tag{6}$$

i.e, the ratio of the minimum cost to the observed.

3. Shifted negative half-normal or exponential distribution

With half-normal or exponential distributions for inefficiency and normally distributed noise, both the expected value and skewness for the composite error become positive. However, in practice, the residuals from the first step in the StoNED procedure does not always have this property. In order to allow for a positive expectation and a negative skewness, simultaneously, we propose to use a shifted negative half-normal or exponential distribution in this study. The modified model is given by

$$\varepsilon_i = v_i' + u_i', \quad u_i' = A - \omega_i, \tag{7}$$

where ε_i is the residual of company i, and v_i' and u_i' are the noise and inefficiency, respectively. The noise term v_i' is assumed to follow a normal distribution $N(0, \sigma_{v'}^2)$, while the inefficiency term is given by the positive constant A minus the stochastic variable ω_i, which follows a positive distribution. We will consider two different one-sided distributions: The half-normal and the exponential. For the half-normal case, i.e., $\omega_i \sim |N(0, \sigma_\omega^2)|$, we have $Var(\omega_i) = \frac{\pi-2}{\pi}\sigma_\omega^2$ and $E(\omega_i) = \sigma_\omega \sqrt{\frac{2}{\pi}}$, hence the expectation and variance of inefficiency are $E(u_i') = A - \sigma_\omega\sqrt{2/\pi}$ and $Var(u_i') = \frac{\pi-2}{\pi}\sigma_\omega^2$, respectively. We will also study the case where ω_i is assumed to follow an exponential distribution with rate $1/\tau$, which implies that the expectation and the variance of inefficiency will be $\mathrm{E}(u_i') = A - \tau$ and $Var(u_i') = \tau^2$, respectively.

We will use the maximum likelihood method to estimate the parameters A, σ_ω, and σ_v' for the half-normal case, and A, τ and σ_v' for the exponential case. In order to perform maximum likelihood estimation, we need the likelihood function of the composite error term $\varepsilon = v' + A - \omega$, which can be expressed as the sum of the variables $Y = v' + A$ and $-\omega$. Hence, the density function of the composite error ε is given by

$$f_\varepsilon(\varepsilon) = \int_{-\infty}^{+\infty} f_\omega(\omega)\, f_Y(\varepsilon + \omega) d\omega = \int_0^{+\infty} f_\omega(\omega)\, f_Y(\varepsilon + \omega) d\omega. \tag{8}$$

The variable Y follows a normal distribution $\left(A, \sigma_{v'}^2\right)$, i.e., with the probability density function

$$f_Y(y) = \frac{2}{\sqrt{2\pi\sigma_{v'}^2}} \exp\left(-\frac{\mathbf{1}}{\mathbf{2}}\frac{(y-A)^2}{\sigma_{v'}^2}\right). \tag{9}$$

For the case where ω follows a half-normal distribution, we have

$$f_\omega(\omega) = \begin{cases} \frac{2}{\sqrt{2\pi\sigma_\omega^2}} \exp\left(-\mathbf{\frac{1}{2}}\frac{\omega^2}{\sigma_\omega^2}\right), & \text{for } \omega \geq 0, \\ 0, & \text{for } \omega \geq 0, \end{cases} \tag{10}$$

and for the exponential case

$$f_\omega(\omega) = \begin{cases} \frac{1}{\tau} \exp\left(-\frac{\omega}{\tau}\right), & \text{for } \omega \geq 0, \\ 0, & \text{for } \omega \geq 0, \end{cases} \tag{11}$$

Summing the log-densities, $\log f_\varepsilon(\varepsilon)$, over all observations and maximizing, we obtain the maximum likelihood estimators of the parameters, for the half-normal case, A, σ_ω, and $\sigma_{v'}$, and for the exponential case, A, τ and $\sigma_{v'}$. Finally, the cost frontier function is estimated by

$$\hat{C}'(\boldsymbol{y}_i) = \hat{\gamma}_i \cdot \exp\left(-\hat{A} + \hat{\sigma}_\omega \sqrt{\frac{2}{\pi}}\right) \tag{12}$$

or

$$\hat{C}'(\boldsymbol{y}_i) = \hat{\gamma}_i \cdot \exp\left(-\hat{A} + \hat{\tau}\right). \tag{13}$$

4. Monte Carlo simulation and empirical illustration

In this section, a dataset with 123 Norwegian electricity distribution companies for the year 2012 is used to illustrate our approach. The single input is total cost, which includes five elements: operations and maintenance costs, value of cost load (quality cost), thermal power losses, capital depreciation, and return on capital. We specify three variables as outputs: high voltage lines, network stations, and the number of customers. High voltage lines and network stations representing structural and environmental conditions may affect required network size and thereby the cost level of the companies. See Cheng et al (2014) for a more detailed description of the dataset.

Descriptive statistics of the composite errors (residuals) obtained from Stage 1 in StoNED are presented in Table 1. The negative value of the skewness statistic suggests that our alternative model, which allows for negative skewness, is indeed suitable.

Table 1 Descriptive statistics of the composite errors

Variable	Min	Max	SD	Mean	Second moment	Third moment	Skewness
Composite error	-0.4857	0.4014	0.1483	0.0009	0.0218	-0.0004	-0.1243

Given our assumption of negative skewness, we perform a simulation study in order to investigate the sensitivity of the cost estimates to the parametric form of the one-sided distribution assumed for ω. In order to do this, we simulate 1000 replications of our cost data under two different distributional assumptions: half-normal and exponential. Then, for each replication, we apply the estimators given by (12) and (13). A priori, we would expect a better fit when we apply an estimator that is consistent with the data generating process, i.e., that (12) is best when the true distribution is the half-normal, and that (13) is best when the true distribution is the exponential.

The starting point for our simulation is, for each company $i = 1, \dots, 123$, a best-practice cost $C'(\boldsymbol{y}_i) = \hat{\gamma}_i \cdot e^{-0.1}$, i.e., the average-practice estimate from Stage 1 in StoNED improved by an industry efficiency factor of $e^{-0.1}$. Next, we simulate noise and inefficiency values for each company based on the half-normal and the exponential distribution, respectively. We use parameter values $A = 0.2$ and $\sigma_{v'} = 0.01$, and for the one-sided distributions we use values such that $A - E(\omega) = 0.1$, i.e., consistent with the assumed industry inefficiency. For the half-normal case this implies $\sigma_\omega = \frac{0.1}{\sqrt{2/\pi}}$, and for the exponential case it implies $\tau = 0.1$. For each of the 1000 replications we then estimate $\hat{C}'(\boldsymbol{y}_i)$ based on both (12) and (13).

Figure 1 shows the results of the simulation study. The 4 panels show scatterplots of the average cost estimate over the 1000 replications for each firm divided by the "true" cost. In the two diagrams to the left, we estimate with (12), i.e., assuming half-normal inefficiencies, and in the rightmost we estimate with (13), i.e., assuming exponential inefficiencies.

As can be seen from the results, the estimators recapture the true cost quite well, i.e., the average estimated costs are close to the true costs, also for the cases when the estimator assumes the wrong distribution. We note, however, that the average relative cost is, to a small extent, systematically overestimated. The reason for this is a topic for future research.

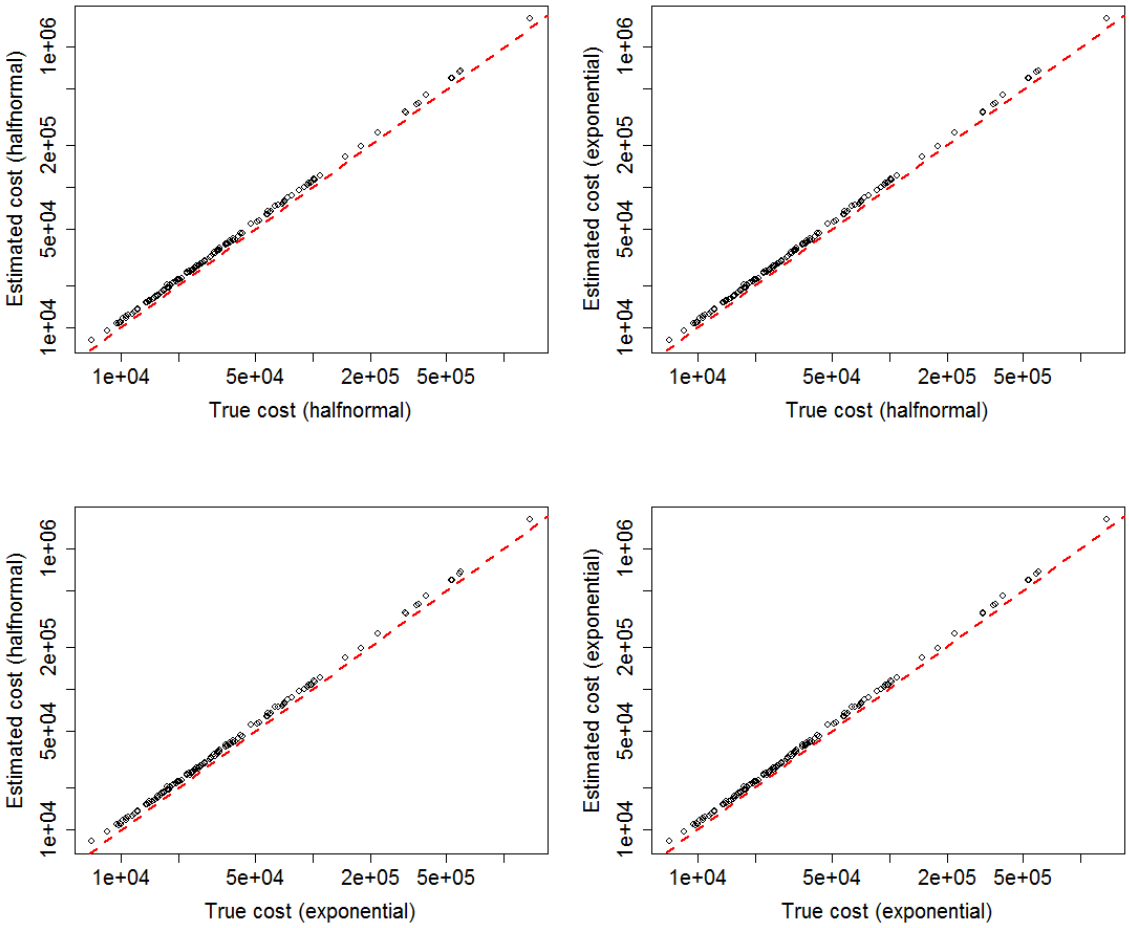

Figure 1 Estimated cost versus true cost

In Figure 2, we show the mean absolute errors between estimated cost and true cost, where the mean is taken over the 1000 replications. Each of the data series shown have 123 data points, one for each of the companies, and we have divided the errors by the true cost for each company. The two diagrams correspond to the respective distributions, i.e., half-normal and exponential, from which we have simulated our cost data. The green circles indicate errors arising when the cost estimation is based on the correct distributional assumption, i.e., consistent with the true distribution, and the red diagonal crosses show what the errors would be if we take the wrong assumption. We observe from Figure 2 that the mean absolute errors are considerable, but that the effect of the distributional assumption is hardly visible. Although the effect is small, it has the expected sign, i.e., the mean absolute errors are always larger when an incorrectly specified distribution for the inefficiency is used.

If an even smaller noise variance, σ_v, is chosen, the distribution of ε will become difficult to distinguish from a normal distribution. None of the two models will then be able to

separate out the effect of inefficiency from that of noise. In fact, we investigated this by changing σ_v to 0.1. For this case, the situation sometimes occur that an incorrectly specified model performs marginally better than a correctly specified one in terms of mean absolute deviations. The term "incorrectly specified", however, is somewhat ambiguous. A better term is perhaps "over-specified". If the data is normally distributed, a model consisting of a normally distributed term will suffice to capture the variation in the data.

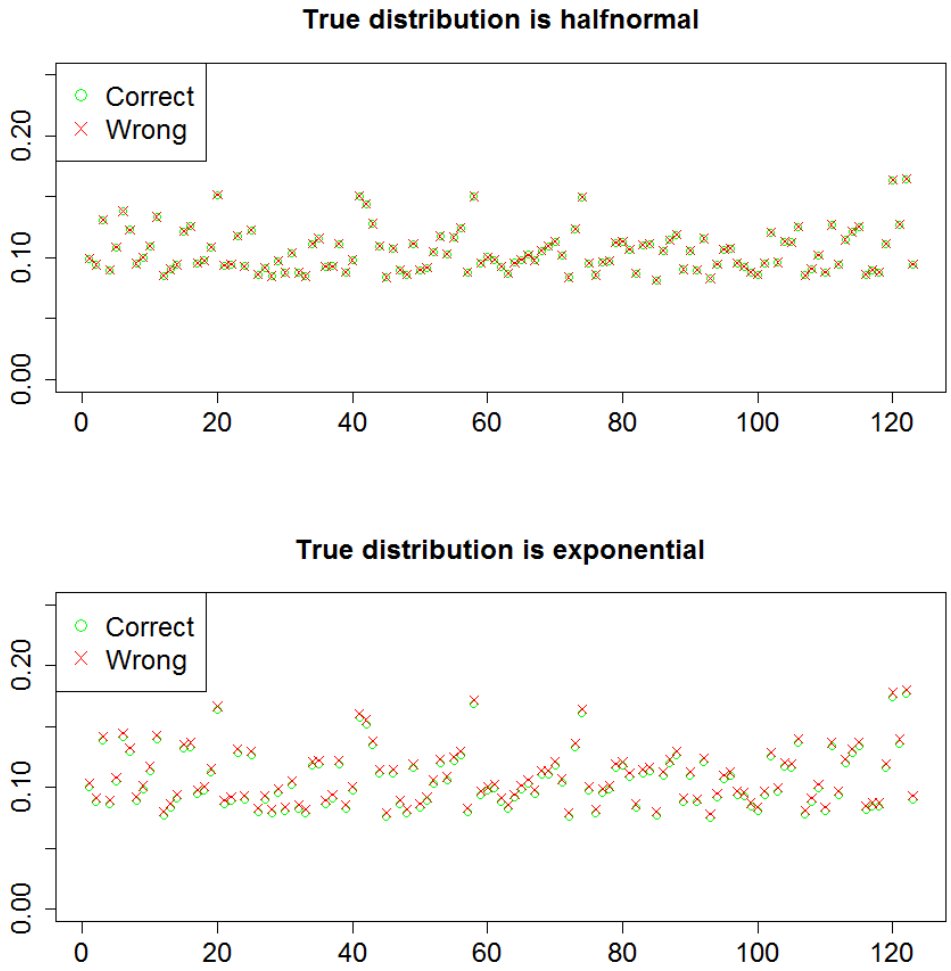

Figure 2 Mean absolute estimation errors, relative to true cost

5. Conclusion

This paper suggest a modification in order to handle situations where the inefficiency has a negatively skewed distribution (but with positive expectation). Furthermore, it studies the impact the distributional assumption has when the cost function is estimated. Our conclusion is that, for the case studied in this paper, the effect is miniscule, compared to other sources of error.

References

Aigner, D. J., Lovell, C. A. K., Schmidt, P., 1977. Formulation and estimation of stochastic frontier models. Journal of Econometrics 6, 21–37.

Almanidis, P., Sickles. R., 2011, Skewness problem in Stochastic Frontier Models: Fact or Fiction? in Exploring Research Frontiers in Contemporary Statistics and Econometrics: A Festschrift in Honor of Leopold Simar. Ingrid Van Keilegom and Paul Wilson (eds.), New York: Springer.

Almanidis, P., Qian, J., Sickles, R., 2011. Bounded Stochastic Frontiers with an Application to the US Banking Industry: 1984-2009, working paper.

Carree, M., 2002. Technological inefficiency and the skewness of the error component in stochastic frontier analysis. Economics Letters 77, 101-107.

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

Cheng, X. M., Bjørndal, E., Bjørndal, M., 2014. Cost efficiency analysis based on the DEA and StoNED models: Case of Norwegian electricity distribution companies. European Energy Market (EEM).

Cornwell, C., Schmidt, P., Sickles, R., 1990. Production Frontiers with Cross Sectional and Time Series Variation in Efficiency Levels. Journal of Econometrics 46:185-200.

Cornwell, C., Schmidt, P., 1996. Production Frontier and Efficiency Measurement, in L. Matyas and P. Sevestre (eds), Econometrics of Panel Data: Handbook of Theory and Application, Kluwer Academic Publishers.

Green, A., Mayes, D., 1991. Technical inefficiency in manufacturing industries. Economic Journal 101, 523–538.

Greene, W., 1980. On the Estimation of a Flexible Frontier Production Model. Journal of Econometrics 3, 101-115.

Johnson, A. L., Kuosmanen, T., 2011. One-stage estimation of the effects of operational conditions and practices on productive performance: asymptotically normal and efficient, root-n consistent StoNED method. Journal of Productivity Analysis 36, 219–230.

Kumbhakar, S., 1990. Production Frontiers and Panel Data, and Time Varying Technical Inefficiency. Journal of Econometrics 46: 201-11.

Kumbhakar, S., 1991. Estimation of technical ineffiency in panel data models with firm and time specific effects. Economic Letters, 36, 43-48.

Kuosmanen, T., 2012. Stochastic semi-nonparametric frontier estimation of electricity distribution networks: Application of the StoNED method in the Finnish regulatory model. Energy Economics 34 (6), 2189–2199

Kuosmanen, T., Kortelainen, M., 2012. Stochastic non-smooth envelopment of data: semi-parametric frontier estimation subject to shape constraints. Journal of Productivity Analysis 38 (1), 11–28.

Meeusen, W., van den Broeck, J., 1977. Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error. International Economic Review18, 435—444.

Olson, J. A., Schmidt, P., Waldman, D. M., 1980. A Monte Carlo study of estimators of stochastic frontier production functions. Journal of Econometrics13 (1):67 -82.

Bogetoft, P., Otto, L., 2011. Benchmarking with DEA, SFA, and R. International Series in Operations Research and Management Science. 157. Springerhttp://dx.doi.org/10.1007/978-1-4419–7961-2_10.

Qu, F., Horrace, W., Wu, G. L., 2012. Wrong Skewness and Finite Sample Correction in Parametric Stochastic Frontier Models. Mimeo.

Schmidt, P., Sickles, R., 1984. Production Frontiers with Panel Data. Journal of Business and Economic Statistics 2: 367-74.

Simar, L., Wilson, P. W., 2010. Inference from cross-sectional, stochastic frontier models. Economic Review 29(1):62–98.

Stevenson, R., 1980. Likelihood Functions for Generalized Stochastic Frontier Estimation. Journal of Econometrics13, 58—66.