ECONOMIC DISCUSSION PAPERS

Efficiency Series Paper 2/2019

The Econometric Measurement of Firms' Efficiency

Luis Orea

Departamento de Economía

Universidad de Oviedo

Available online at: **http://economia.uniovi.es/investigacion/papers**

The Econometric Measurement of Firms' Efficiency

Luis Orea [a,b]
Oviedo Efficiency Group
Department of Economics
University of Oviedo

Abstract

This working paper serves as guide to efficiency evaluation from an econometric perspective. The analytical framework relies on the most general parametric models and up to date representations of the production technology through Translog and Quadratic distance functions. We outline the most popular estimation methods: maximum likelihood, method-of-moments and distribution-free approaches. In the last section we discuss more advance topics such as how to control for observed and unobserved environmental variables or endogeneity issues. Other topics examined are dynamic efficiency measurement, production risk and uncertainty, and the decomposition of Malmquist productivity indices.

Keywords: Production and, distance functions, stochastic frontier analysis, environmental variables, endogeneity, dynamic behaviour, productivity growth.

JEL codes: C4, C5, C6, D24.

[a] Part of the content of this paper will be published as a chapter in the forthcoming Springer's book *Data Science and Productivity Analytics*.

[b] Luis Orea would like to thank the financial support obtained from the Government of the Principality of Asturias and the European Regional Development Fund (ERDF). This working paper was also supported by the Spanish Ministry of Economics, Industry and Competitiveness (Project MINECO-18-ECO2017-85788-R).

1. Introduction

In this paper we summarize the main features of the standard econometric approach to measuring firms' inefficiency (and productivity). Given that the efficient production/cost of each firm is not directly observed, it must be inferred from real data using *frontier* models that involves the estimation of both the technological parameters and the parameters of firms' inefficiency. In this paper we provide guidance on the options that are available in order to successfully undertake research in this field using the so-called *Stochastic Frontier Analysis* (SFA) models, the most popular parametric frontier technique.[1] This ranges from the selection of the appropriate theoretical model to the use of the empirical techniques best suited to achieving reliable results. The selection of analytical frameworks and methods presented in the paper is necessarily partial, as it is virtually impossible to cover all recent research in such a dynamic area.[2]

We start this paper summarizing in Section 2 the main results of production theory; particularly the possibility of characterizing the behaviour of the firm from the primal−technological perspective. As firms produce multiple outputs using multiple inputs, the primal representation of the technology relies on the concept of distance function, which is also interpreted as a measure of productive performance. We discuss in this section the choice of functional forms when representing firms' technology and examine the advantages and drawbacks of the so-called *flexible* functional forms.

Section 3 outlines the most popular estimation methods available to undertake SFA efficiency analyses. The most popular estimation method is *maximum likelihood*, where the parameters of the distance (production) function and the random term capturing firms' inefficiency are estimated in a single stage. A second method is the *method-of-moments* approach, where the distance function is first estimated using standard econometric techniques and distributional assumptions are only invoked in a second stage to estimate the parameter(s) describing the structure of the error components. Unlike the two abovementioned methods, firms' efficiency scores can also be computed without making specific distributional assumptions using the so-called *distribution-free approach*.

Section 4 discusses more advance topics and extends somehow the basic models introduced in the previous section. In Subsection 4.1, we examine how to control for environmental or contextual variables that do not fall within managerial discretion. To deal with this issue, non-discretionary variables can be included not only as frontier regressors but also as determinants of firms' inefficiency, and we discuss the implications of several empirical strategies to achieve this aim.

Subsection 4.2 presents a series of recent models addressing endogeneity issues using the SFA approach. Some focus on the correlation between the regressors and the noise term, while others address the correlation with the inefficiency term. Models can be estimated using different techniques and using one or two-stage methods. In Subsection 4.3 we outline several frontier models that control for *unobserved* differences in firms' technology or environmental conditions. In particular, we examine several panel-data models introduced by Greene (2005)

[1] Another popular but non-parametric frontier technique is mathematical programming or Data Envelopment Analysis (DEA). The individual results from parametric and non-parametric methods will generally differ. However, the difference between the two methods is less pronounced nowadays than they used to be because both approaches now benefit from recent advances that address their shortcomings.

[2] For a comprehensive survey of this literature, we recommend the following references: Kumbhakar and Lovell (2000), Parmeter and Kumbhakar (2014), and Kumbhakar et al. (2015).

and extended later on by other researchers; the latent class stochastic frontier models, and the recent spatial frontier models that takes into account the spatial structure of the data.

Subsection 4.4 is devoted to dynamic efficiency measurement. Considering as a departure point the existence of rigidities associated to fixed inputs, information failures when planning investment decisions, etc., dynamic modelling emerges naturally. Here we discuss two main approaches by which to incorporate the dynamic nature of the decision-making process into efficiency analyses. One approach is to use reduced-form models that do not require explicit modelling of the firm's dynamic behaviour, which in turn do not impose strong assumptions on the data. The second approach makes use of structural models that make explicit assumptions with respect to the firm's economic objectives.

In Subsection 4.5 we summarize several approaches proposed in the applied literature to account for production risk, stochastic technologies and uncertainty. Indeed, it has been found that ignoring the stochastic nature of firms' performance may have relevant welfare and policy implications, as the latter would be based on biased estimates and misleading inference.

Finally, we study in Subsection 4.6 the popular Malmquist productivity index and its decomposition into several terms explaining productivity change (efficiency change, technical change…) based on estimated production and distance functions. The decomposition is built upon by relying on economic theory approach to index numbers and the exactness between the former Malmquist productivity index and some of the flexible functional forms presented in our theoretical background section.

2. Theoretical background

The point of departure of any theoretical and empirical study of efficiency and productivity is whether it is merely concerned with technical performance from an engineering perspective or it has an economic dimension. The technical or engineering approach is the only available choice when prices are unavailable (for example, the public sector provision of some public goods and services), or when they simply do not exist (for example, undesirable by-products such as waste and pollution). In this case we presume that the objective of the firm is technological, based on quantities only, and technology must be inferred using a primal approach, such as production or distance functions. On the contrary, as would be the case of firms in an industry, if market prices for inputs and outputs are available, then we can extend our engineering analysis to the firm's market environment. In this case we presume that the objective of the firm is economic, and its analysis requires data on quantities and prices and dual representations of firms' technology such as cost and profit functions.

2.1 Primal approach: production and distance functions

2.1.1. Definitions and properties

For the single output case, the technology can be represented by the production function defined as the maximum amount of output that can be obtained from any combination of inputs:

$$f(x) = max\{y: (x, y) \in T\} \quad (1)$$

where T is the technology set. In the multi-output case, a suitable representation of the technology is given by the distance function introduced by Shephard (1970). This representation can be made from alternative orientations including the following output and input-oriented distance functions:

$$D_O(x, y) = min\{\theta: (x, y/\theta) \in T\} \quad (2)$$

$$D_I(x, y) = max\{\lambda: (x/\lambda, y) \in T\} \quad (3)$$

If the technology satisfies the customary axioms, the *output* distance function, ODF, has the range $0 \leq D_O(x,y) \leq 1$. It is homogeneous of degree one in outputs, non-decreasing in outputs and non-increasing in inputs. Notice that the advantage of this interpretation is that it leaves room for technical inefficiency when $D_O(x,y) < 1$. In this case, value of the output distance function can be directly interpreted as a measure of firms' technical efficiency, that is $ET = D_O$. In contrast, the *input* distance function, IDF, has the range $D_I(x,y) \geq 1$. It is homogeneous of degree one in inputs, non-decreasing in inputs, and non-increasing in outputs. A firm is inefficient when $D_I(x,y) > 1$. Therefore, firms' technical efficiency can be measured as $ET = 1/D_I$.[3]

More recent and flexible characterizations are the additive *directional* distance functions that can be defined as:

$$\vec{D}(x,y,-g_x,g_y) = max\{\tau: (x - \tau g_x, y + \tau g_y) \in T\} \quad (4)$$

The directional distance function, DDF, measures the simultaneous maximum reduction in inputs and expansion in outputs given a pre-specified directional vector defined by $g = (g_x, g_y)$ and the actual technology. The properties of this function are presented in Chambers et al. (1996, 1998). Just mention here that this function nests Shephard's input and output distance functions depending on the specific values of the directional vector.

2.1.2. The importance of imposing theoretical properties.

Notice that, at first sight, the distance functions in (2)-(4) dependent on the same vector of inputs and outputs. Thus, if we were able to estimate a function of inputs and outputs, say $D(x,y)$, how do we ensure that we have estimated our preferred choice, say, an output distance function, and not an input distance function? For identification purposes we need to take advantage of one of the properties of distance functions. In particular, the key property for identification is the homogeneity condition for the input and output distance functions and the translation property for the directional distance functions. The latter property is the additive analogy to the multiplicative homogeneity property of Shephard's distance functions. Identification works because each homogeneity condition involves different sets of variables.[4]

For instance, in an output distance function, the linear homogeneity condition of the distance function implies that $D(x,\lambda y) = \lambda D(x,y)$. If we assume that $\lambda = 1/y_M$, we get after taking logs that:

$$lnD = lnD(x, y/y_M) + lny_M \quad (5)$$

The term measuring firms' inefficiency (i.e. lnD) is not observed by the researcher and thus it cannot be used as a proper dependent variable to estimate. However, the linear homogeneity condition immediately "produces" an observed dependent variable for the above model if we rewrite (5) as:

$$-lny_M = lnD(x, y/y_M) + u \quad (6)$$

where $u = -lnD$, or

[3] I have often been asked whether it is possible to compute the elasticity of a specific input (output) variable using distance functions due to its radial definition involving a whole set of inputs (outputs). The answer to this question is: yes. In the Appendix we show how to compute relevant economic properties of the multi-input multi-output distance function such as specific input and output elasticities or marginal effects, and the scale elasticity regardless we use an input or output-oriented distance functions.

[4] Interestingly, although the underlying technology is the same, the coefficients of each distance function differ.

$$lny_M = -lnD(x, y/y_M) - u \quad (7)$$

Note that this ODF collapses to a standard production function if M=1, and that we have reversed the signs of all the coefficients of $lnD(\cdot)$. Therefore, the estimated parameters can be interpreted as the coefficients of a (multi-output) production function. A similar expression can be obtained if we impose the linear homogeneity condition in inputs rather than in outputs, and an input distance function is estimated instead.[5]

The choice of orientation should be determined, at least partially, by the capability of firms to adjust their inputs and outputs in order to become fully efficient. However, Kumbhakar et al. (2007) show that, once the distance function is known, input (output) oriented inefficiency scores can be obtained from output (input) distance functions. To see this clearly, assume that we want to estimate the output distance function (5) but using an input-oriented measure of firms' efficiency. The equation to be estimated can be written as:

$$0 = \ln D(xe^{-\eta}, y/y_M) + lny_M \quad (8)$$

where now η measures firms' efficiency in terms of input reductions, conditional on the observed output vector. Thus, if any measure of firms' inefficiency can be estimated using any primal representation of firms' technology, why is the choice of orientation a relevant issue? It is a relevant issue for at least two empirical reasons. First of all, because both the efficiency scores and the estimated technologies are expected to be different due to neglected endogeneity issues. The choice of orientation is also relevant for the "complexity" of the stochastic part of the model in a SFA model. For instance, Kumbhakar and Tsionas (2006) show that the standard maximum likelihood (ML) method cannot be applied to estimate input-oriented production functions. This issue is examined later once several functional forms for the distance functions have been introduced.

Regarding the directional distance function, while its general specification is given in (4), quite often the directional vector is set to $(g_x, g_y) = (1,1)$. In this case, this function can be written as:

$$\vec{D} = \vec{D}(x, y; -1,1) = \vec{D}(x, y) \quad (9)$$

If the above directional distance function satisfies the *translation property* that says that if output is expanded by α and input is contracted by α, then the resulting value of the distance function is reduced by α:

$$\vec{D}(x - \alpha,\ y + \alpha,\ ) = \vec{D}(y, x) - \alpha \quad (10)$$

Thus, replacing above $\vec{D}(x, y)$ with $\vec{D}(x - \alpha, y + \alpha\ ) + \alpha$, we get:

$$-\alpha = \vec{D}(x - \alpha, y + \alpha,\ ) - u \quad (11)$$

where now $u = \vec{D}$. We obtain variation on the left-hand side by choosing an α that is specific to each firm. For instance, $\alpha = y_M$.

2.1.3. Functional forms

The initial and most commonly employed distance functions (or, equivalently, their corresponding production functions in the single output case), i.e., Cobb-Douglas (CD) or

[5] The linear homogeneity condition in inputs yields the following IDF:

$$lnx_J = f(x/x_J, y) + u$$

where now $u = lnD \geq 0$ measures firms' inefficiency in terms of inputs, and $f(x/x_J, y)$ is non-increasing in inputs, and non-decreasing in outputs. Therefore $f(x/x_J, y)$ can be interpreted as an input requirement function.

Constant Elasticity of Substitution (CES), as well as their associated dual functions, place significant restrictions on technological and economic behaviour relations. For example, in production analysis they restrict all output and input elasticities to be common to all firms and returns to scale do not vary with firms' size; while for cost minimization, the linear or log-linear specifications imply that inputs demand, or the share of each input in costs, are independent of the output level.[6] While these characteristics are quite restrictive, these functions are "well-behaved" and satisfy all desirable neoclassical properties, particularly they are continuous and twice differentiable. In turn, this ensures that relevant theoretical results based on the envelopment theorem -i.e., Shephard and Hotelling's lemma, allow the recovery of the demand and supply equations- without solving their primal functions, and that comparative statics exercises can be easily performed.

A subsequent generation of technological representations beyond the CES production function nesting the CD, linear and fix-proportions technologies, emerged in the 70s with the so-called *second order flexible functional forms* that permit a more general representation of the production technology (see Diewert, 1971; p. 481-507). The specifications can be seen as second order Taylor-series mathematical expansions around different points with different transformations of the variables -e.g., quadratic, Leontief, or Translog, while successive functional forms are based on higher order Laurent and Fourier expansions (Thompson, 1988). One advantage of the latter proposal is that it provides a global rather than a local approximation to the underlying technology; but since its econometric estimation and parameter interpretation prove more demanding, they are by far less popular in empirical research.

The fact that the number of parameters to be estimated increases exponentially with the number of variables included in the functional form, empirical research is *de facto* restricted to the quadratic approximation. If a large sample cannot be collected, degrees of freedom can be easily exhausted, and a general practice is to aggregate commodities and prices; but consistent aggregation is only possible under strong restrictions on the underlying technology—e.g., separability. The properties of flexible functional forms ultimately determine whether they are globally well-behaved in the presence of large data variability. For instance, the Quadratic specification fails to satisfy the regularity conditions over the entire range of sample observations. However, how to test those global properties and impose regularity conditions globally remains unclear because imposing regularity conditions globally often comes at the cost of limiting the flexibility of the functional form. Given this trade-off, the common practice is to evaluate the estimated functions at the sample mean, rather than at each individual observation.[7]

Despite these caveats, flexible functional forms are useful and have become standard in empirical studies. To exemplify their capabilities when testing functional, we show two representative specifications. The first one makes use of the *Translog* formulation to specify the output distance function, and the second one corresponds to the *Quadratic* directional distance function.

As for the Translog output distance function with output-oriented inefficiency, the specification corresponds to:

[6] The limitations of the Cobb-Douglas functions when testing the neoclassical theory of the firm constituted the basis for newer, less restrictive functional forms (Zellner and Revankar, 1969).

[7] It should be pointed out, however, that it is possible to maintain *local* flexibility using Bayesian techniques. See Griffiths et al. (2000) and O'Donnell and Coelli (2005).

$$\ln y_M = \beta_0 + \sum_{j=1}^{J} \beta_j \ln x_j + \sum_{m=1}^{M-1} \beta_m \ln y_m^* + \frac{1}{2}\sum_{j=1}^{J} \beta_{jj} \ln x_j{}^2 + \frac{1}{2}\sum_{m=1}^{M-1} \beta_{mm} \ln y_m^{*\,2} + \sum_{j=1}^{J}\sum_{k\neq j}^{K} \beta_{jk} \ln x_j \ln x_k + \frac{1}{2}\sum_{m=1}^{M-1}\sum_{n\neq m}^{N} \beta_{mn} \ln y_m^* \ln y_n^* + \sum_{m=1}^{M-1}\sum_{j=1}^{J} \beta_{mj} \ln y_m^* \ln x_j - u \quad (12)$$

where $\ln y_m^* = \ln y_m - \ln y_M$. Note that the output-oriented inefficiency term appears above as an additive term. Therefore, the above parameters can be easily estimated using the standard maximum likelihood (ML) techniques because the typical distributional assumptions for u provide a closed-form for the distribution of the error term. If instead we are willing to a Translog output distance function using an input-oriented measure of firms' efficiency, the model to be estimated is:

$$\ln y_M = \beta_0 + \sum_{j=1}^{J} \beta_j \ln\left(x_j e^{-\eta}\right) + \sum_{m=1}^{M-1} \beta_m \ln y_m^* + \frac{1}{2}\sum_{j=1}^{J} \beta_{jj} \ln\left(x_j e^{-\eta}\right)^2 + \frac{1}{2}\sum_{m=1}^{M-1} \beta_{mm} \ln y_m^{*\,2} + \sum_{j=1}^{J}\sum_{k\neq j}^{K} \beta_{jk} \ln\left(x_j e^{-\eta}\right) \ln\left(x_k e^{-\eta}\right) + \frac{1}{2}\sum_{m=1}^{M-1}\sum_{n\neq m}^{N} \beta_{mn} \ln y_m^* \ln y_n^* + \sum_{m=1}^{M-1}\sum_{j=1}^{J} \beta_{mj} \ln y_m^* \ln\left(x_j e^{-\eta}\right) \quad (13)$$

Assuming one input, the model can be written as:

$$\ln y_M = -lnD(x, y/y_M) - \left[\beta_j + \beta_{jj} + \sum_{m=1}^{M-1} \beta_{mj} \ln y_m^*\right]\eta + \beta_{jj}\eta^2 \quad (14)$$

The presence of the η^2 term makes the derivation of a closed likelihood function impossible, and this precludes using standard ML techniques. Similar comments can be made if we were to use a directional distance function. In all cases where we have intractable likelihood functions, they can be maximized by simulated maximum likelihood.[8] A final important remark regarding equation (14) is that the output orientation of the distance function does not force the researcher to use an input-oriented measure of firms' inefficiency. We first do it just for simplicity, and in doing so are likely to attenuate endogeneity problems as well.

As for the directional distance function, the reason why the quadratic formulation is the best choice is that the translation property can be easily imposed on this specification—just as the homogeneity properties corresponding to the radial input or output distance functions can be easily imposed on the Translog specification. Once the translation property is imposed using $\alpha = y_M$, the quadratic specification of (11) can be written as:

$$-y_M = \beta_0 + \sum_{j=1}^{J} \beta_j x_j^* + \frac{1}{2}\sum_{j=1}^{J}\sum_{k=1}^{K} \beta_{jk}\, x_j^* x_k^* + \sum_{m=1}^{M-1} \beta_m y_m^* + \frac{1}{2}\sum_{m=1}^{M-1}\sum_{n=1}^{N} \beta_{mn} y_m^* y_n^* + \sum_{m=1}^{M-1}\sum_{j=1}^{J} \beta_{mj} y_m^* x_j^* - u \quad (15)$$

where $x_j^* = x_j - y_M$ and $y_m^* = y_m + y_M$. It is worth mentioning that inefficiency is measured here in physical units, and not in percentage terms as it happens is we use a traditional Translog specification. Both measures are correct, albeit they are simply using different approaches to measure the distance to the frontier. On the other hand, an interesting feature that is often overlooked is that the Quadratic specification is normally estimated once the variables are normalized with the sample means (see Färe et al. 2005; p. 480). As the normalized variables are unit free, in practice the estimated inefficiency scores can be interpreted as proportional changes in outputs and inputs, in the same fashion as in the standard radial distance functions.

[8] As shown by Parmeter and Kumbhakar (2014; p. 52) using a Translog cost function, if the production technology is homogeneous in outputs, the model can be estimated using simple ML techniques.

2.2 Dual approach: cost functions

We introduce here a cost minimization objective in order to discuss the duality framework allowing for an overall economic efficiency analysis. Based on the previous primal representations of the technology, and considering the vectors of input prices, w, the following cost function can be defined:

$$C(y,w) = \min_{x}\{wx: (x,y) \in T\} \tag{16}$$

The cost function represents the minimum cost of producing a given amount of outputs, and assuming the necessary derivative properties—including continuity and differentiability, yields the input demand functions by applying Shephard's lemma.[9] If the technology satisfies the customary axioms, the cost function (16) is homogeneous of degree one in input prices, and non-decreasing in outputs and in input prices. Chambers et al. (1998) prove the duality between the input distance functions and its associated cost function. Unlike the distance function that only provides a measure of technical efficiency, the above definition leaves room for both technical and allocative inefficiency. However, Kumbhakar et al. (2015) point out that outputs and input prices are endogenous if firms are allocative inefficient because in this case the traditional u term depends on y and w.

Regarding the functional forms, the Translog cost function corresponds to:

$$\begin{aligned} ln\left(\frac{C}{w_J}\right) = \beta_0 + \sum_{j=1}^{J-1}\beta_j \ln\left(\frac{w_j}{w_J}\right) + \sum_{m=1}^{M}\beta_m \ln y_m + \frac{1}{2}\sum_{j=1}^{J-1}\beta_{jj}\ln\left(\frac{w_j}{w_J}\right)^2 + \frac{1}{2}\sum_{m=1}^{M}\beta_{mm}\ln y_m^2 \\ + \sum_{j=1}^{J-1}\sum_{k\neq j}^{K-1}\beta_{jk}\ln\left(\frac{w_j}{w_J}\right)\ln\left(\frac{w_k}{w_J}\right) + \sum_{m=1}^{M}\sum_{n\neq m}^{N}\beta_{mn}\ln y_m \ln y_n \\ + \sum_{m=1}^{M}\sum_{j=1}^{J-1}\beta_{mj}\ln y_m \ln\left(\frac{w_j}{w_J}\right) + u \end{aligned} \tag{17}$$

where u measures firms' technical and allocative inefficiency in terms of cost increases. Notice that we have already imposed linear homogeneity in input prices in the above cost function, and that the input-oriented inefficiency term appears above as an additive term. Again, this implies that the above parameters can be estimated by ML. Applying the Shephard's lemma in (17), we get the following cost share equations:

$$\begin{aligned} S_1 &= \beta_1 + \beta_{11}\ln\left(\frac{w_1}{w_J}\right) + \sum_{k\neq 1}^{J-1}\beta_{1k}\ln\left(\frac{w_k}{w_J}\right) + \sum_{m=1}^{M}\beta_{m1}\ln y_m \\ &\vdots \\ S_{J-1} &= \beta_{J-1} + \beta_{J-1,J-1}\ln\left(\frac{w_{J-1}}{w_J}\right) + \sum_{k\neq J-1}^{J-2}\beta_{J-1k}\ln\left(\frac{w_k}{w_J}\right) + \sum_{m=1}^{M}\beta_{mJ-1}\ln y_m \end{aligned} \tag{18}$$

In principle, estimating the cost system (17)-(18) is more efficient from a statistical perspective because no additional parameters are added to the model.[10] However, Kumbhakar et al (2015) clearly show that estimating a cost system using (17) and (18) is problematic as, except with input-oriented technical inefficiency and zero allocative inefficiency. For this

[9] It is also possible to define shadow cost functions $C(y,w^s)$ constituting the dual representation of the technology for non-market oriented (i.e., non-profit) organizations (e.g., public goods such as the provision of health and education services). In this case, for instance, the so-called shadow prices w^s rationalize the observed input quantity vector *x* as a cost-minimizing choice for the observed output vector *y*. If the minimum-cost condition is satisfied, the shadow price vector equals the market price vector. Rodríguez-Álvarez and Lovell (2004) show that these vectors may differ as a result of utility maximizing behavior on the part the bureaucrat, restricted by a budget constraint.

[10] This happens if we do not allow for non-zero mean values of the error terms traditionally added to each cost share equation in (18).

reason, they strongly prefer estimating *primal* system of equations consisting of a stochastic production (distance) function and a set of first order conditions for cost minimization.

3. Estimation methods

In this section we outline the most popular parametric frontier techniques aiming to measure both firms' inefficiency and technology. For notational ease, we develop this and next sections for cross-sectional data, except when it is compulsory to use a panel data framework. We also confine our discussion to the estimation of technical efficiency using output distance functions because they can be interpreted as a traditional but multi-output production function.[11] Thus, firm performance is evaluated by means of the following distance function:

$$lny_{Mi} = -\ln D\left(x_i, \frac{y_i}{y_{Mi}}, \beta\right) + v_i - u_i \tag{19}$$

where the subscript *i* stands for firm, β is now a vector of technological parameters, v_i is a two-sided noise term, and $u_i = -lnD_i \geq 0$ is a one-sided term capturing firms' inefficiency. In equation (19) we specify the distance function as being stochastic in order to capture random shocks that are not under the control of the firm. It can also be interpreted as a specification error term that appears when the researcher tries to model the firm's technology. Note also that this model can be immediately estimated econometrically once a particular functional form is chosen for $\ln D(x_i, y_i/y_{Mi}, \beta)$, and u_i is properly modelled.[12]

Note also that the composed error term $\varepsilon_i = v_i - u_i$ in (19) comprises two independent parts, a noise term and an inefficiency term. They are likely to follow different distributions given their different nature. Indeed, it is conventionally assumed that $v_i$follows a symmetric distribution since random shocks and specification errors might take both positive and negative values. However, by construction, inefficient performance always produces a contraction in firms' output. For this reason, u_i is assumed to be non-negative (and asymmetrically) distributed. This results in a composed error term ε_i that is asymmetrically distributed. As customary in the literature, it is also assumed throughout that both random terms are distributed independently of each other and of the input variable.

We now turn to explaining how to estimate the above frontier model. The estimation of the model involves both the parameters of the distance (production) function and the inefficiency. Even with very simple SFA models, the researcher has several estimation methods at hand, and, in most applications, chooses only one. All have their own advantages and disadvantages. Equation (19) can first be estimated via *maximum likelihood* (ML) once particular distributional assumptions on both random terms are made. ML is the most popular empirical strategy in the literature, but it relies on (perhaps strong) assumptions regarding the distribution of these terms, and the exogenous nature of the regressors. Both technological parameters of the distance function (β) and the structure of the two error components (i.e., the variance of v_i and u_i) are estimated simultaneously in a single stage using ML. In this sense, ML merges the two stages of the following estimation method.

A second method that we can choose is the *method-of-moments* (MM) approach, where all technological parameters of the distance function are first estimated using standard econometric techniques (e.g. OLS, IV or GMM) without making specific distributional assumptions on the error components. This stage is independent of distributional assumptions

[11] Although most early SFA applications used production functions, the distance function became as popular as the production functions since Coelli and Perelman (1996), who helped practitioners to estimate and interpret properly the distance functions.

[12] The input distance functions as well as the directional distance function deserve similar comments.

in respect of either error component. In the second stage of the estimation procedure, distributional assumptions are invoked to obtain ML estimates of the parameter(s) describing the structure of the two error components, conditional on the first-stage estimated parameters.[13] Although the MM approach is much less used by practitioners than the traditional ML approach, the most comprehensive SFA versions of the MM estimator are becoming increasingly popular among researchers because it allows for instance dealing with endogenous variables (see Guan et al, 2009), or distinguishing between transient and permanent efficiency (Filippini and Greene, 2016).

Once the model has been estimated using either ML or MM, the next step is to obtain the efficiency values for each firm. They are often estimated by decomposing the estimated residuals of the production function. Following Jondrow et al. (1982), both the mean and the mode of the conditional distribution of u_i given the composed error term ε_i can be used as a point estimate of u_i.

Firms' efficiency scores can also be computed without making specific distributional assumptions on the error components using the so-called *distribution-free approach*. This approach includes the well-known COLS method for cross-sectional data, and the SS and CSS methods for panel data settings. As Kumbhakar et al. (2015; p. 49) remark, the drawback of this approach is that the statistical properties of the estimator of u_i may not be readily available.

3.1. ML estimation

3.1.1. Single equation models

In order to estimate equation (19) using ML, we are forced to choose a distribution for v_i and u_i. The noise term is often assumed to be normally distributed with zero mean and constant standard deviation, i.e. $v_i \sim N(0, \sigma_v)$, with the following density function:

$$f(v_i) = \phi\left(\frac{v_i}{\sigma_v}\right) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{v_i^2}{2\sigma_v^2}\right) \tag{20}$$

Note that $v_i = \varepsilon_i + u_i$, where $\varepsilon_i = lny_{Mi} - X'_{it}\beta$, and $X'_{it}\beta$ is the log of the frontier production (distance) function (e.g., Translog). So,

$$f(v_i) = f(\varepsilon_i + u_i) = \frac{1}{\sqrt{2\pi}\sigma_v} \exp\left(-\frac{(\varepsilon_i+u_i)^2}{2\sigma_v^2}\right) \tag{21}$$

Regarding the inefficiency term, several distributions have been proposed in the literature for this one-sided random term, viz., half-normal (Aigner et al., 1977), exponential (Meeusen and van den Broeck, 1977), and gamma (Greene, 1990). By far, the most popular distribution is the half-normal, which is the truncation (at zero) of a normally-distributed random variable with zero mean and constant standard deviation, that is $u_i \sim N^+(0, \sigma_u)$.[14] Note that, for notational ease, we use σ_u to indicate hereafter the standard deviation of the pre-truncated normal distribution, and not the standard deviation of the post-truncated variable. If u_i follows a (homoscedastic) half-normal distribution, its density function can be written as:

[13] Both variances can also be estimated using the second and third moments of the composed error term taking advantage of the fact that, while the second moment provides information about both variances, the third moment only provides information about the asymmetric random conduct term.

[14] The most important characteristic of this distribution is that the modal value of the inefficiency term is close to zero, and higher values of u_i are increasingly less likely. Stevenson (1980) relaxed the somehow strong assumption that the most probable value is being fully efficient by introducing the truncated-normal distribution, which replaces the zero mean of the pre-truncated normal distribution by a new parameter to be estimated.

$$f(u_i) = \frac{2}{\sigma_u}\phi(\frac{u_i}{\sigma_u}) = \frac{2}{\sqrt{2\pi}\cdot\sigma_u}\exp\left(-\frac{u_i^2}{2\sigma_u^2}\right) \quad (22)$$

Assuming that of v_i and u_i are distributed independently, the density function of the composed error term $\varepsilon_i = v_i - u_i$ can be written as:

$$f(\varepsilon_i) = \int_0^\infty f(\varepsilon_i + u_i)\cdot f(u_i)du_i \quad (23)$$

Given the assumed distributions, the above integration can be computed analytically. The density function of the composed error term of a normal-half-normal model is:

$$f(\varepsilon_i) = \frac{2}{\sigma}\phi\left(\frac{\varepsilon_i}{\sigma}\right)\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right) = \frac{2}{\sigma}\left[\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{\varepsilon_i^2}{2\sigma^2}\right)\right]\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right) \quad (24)$$

where $\sigma = \sqrt{\sigma_u^2 + \sigma_v^2}$ and $\lambda = \sigma_u/\sigma_v$. Therefore, the log-likelihood function for the whole sample (assuming N observations) can be written as:

$$lnLF = \frac{N}{2}\ln(2/\pi) - Nln\sigma - \sum_{i=1}^N \frac{\varepsilon_i^2}{2\sigma^2} + \sum_{i=1}^N ln\Phi\left(-\frac{\varepsilon_i\lambda}{\sigma}\right) \quad (25)$$

Notice that the standard distributional assumptions for v_i and u_i provide a closed-form for the distribution of the composed error term, making the direct application of ML straightforward. The model is simply estimated by choosing the parameters that maximize the likelihood function (25). Newer models are appearing in the literature that do not yield tractable likelihood functions and must be estimated by simulated maximum likelihood. See Parmeter and Kumbhakar (2014, section 7) for an excellent review of recent contributions dealing with this issue. To catch an idea about how this approach works, let us point out that the model can be estimated if we integrate out u_i from $f(\varepsilon_i + u_i)$ in (21):

$$f(\varepsilon_i) = \int_0^\infty \frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{(\varepsilon_i+u_i)^2}{2\sigma_v^2}\right)f(u_i)du_i \quad (26)$$

Notice that the integral can be viewed as an expectation, which we can evaluate through simulation as opposed to analytically. Taking many draws, the above integral can be approximated as:

$$f(\varepsilon_i) \approx \frac{1}{R}\sum_{r=1}^R \frac{1}{\sqrt{2\pi}\sigma_v}\exp\left(-\frac{(\varepsilon_i+\sigma_u|U_r|)^2}{2\sigma_v^2}\right) \quad (27)$$

The final task is obtaining the efficiency scores for each firm. As the procedure to get these scores is the same in ML and MM, it is explained later on in Subsection 3.2.

3.1.2. System models

The previous discussion is concerned with the technical side of the firm. The allocation of inputs in a production model, or in our output distance function, is assumed to be either 100% efficient or they are assumed to be exogenously given. Recent developments in duality theory allow the decomposition of overall economic efficiency into technical and allocative terms in a consistent way. When the production or cost function and the structure of the two error components should be estimated, Parmeter and Kumbhakar (2014) summarize the existing methods, favouring those relying on the primal perspective that are easier to identify and estimate, over systems of equations based on the dual cost approach. Their preferred approach estimates a system consisting of a stochastic production (distance) function, which allows for technical inefficiency, and a set of first order conditions (FOC) for cost minimization, which allow for allocative inefficiency if the FOCs are not fulfilled.

Following Kumbhakar et al. (2015, pp. 210-223), this primal system of equations can be written using a two-input Cobb-Douglas distance function as:

$$\ln y_{Mi} = \alpha_0 + \alpha_1 \ln x_{1i} + \alpha_2 \ln x_{2i} + f\left(\frac{y_i}{y_{Mi}}, \beta\right) + v_i - u_i$$

$$\ln(\alpha_2/\alpha_1) - \ln(w_{2i}/w_{1i}) - \ln x_{2i} + \ln x_{1i} = \xi_{2i} \quad (28)$$

where $v_i \sim N(0, \sigma_v)$, $u_{it} \sim N^+(0, \sigma_u)$, and $\xi_i \sim N(\rho, \sigma_\xi)$.[15] The likelihood function of the whole system is:

$$LF_i = g(v_i - u_i) \cdot d(\xi_i) \cdot |J_i| \quad (29)$$

where $g(v_i - u_i)$ is the density function of a normal-half-normal random variable, $d(\xi_i)$ is the probability density function for ξ_i , and $|J_i|$ is the determinant of the Jacobian matrix:

$$|J_i| = \left|\frac{\partial(v_i - u_i, \xi_i)}{\partial(\ln x_{1i}, \ln x_{2i})}\right| \quad (30)$$

After estimating the parameters of the model by ML, firm-specific efficiency scores can be computed using the Jondrow et al. (1982) formula. Allocative inefficiency can be obtained from the residuals of the FOCs. If $\xi_i < 0$, input x_{2i} is overused relative to input x_{1i}, underused otherwise.

3.2. MM estimation

The MM approach involves three stages. In the first stage, we ignore the structure of the composed error term and estimate the frontier parameters using OLS if the explanatory variables are exogeneous or GMM if they are endogenous.[16] Taking expectations in (19), the model to be estimated in the first stage can be written as:

$$lny_{Mi} = E\left(lny_{Mi} \middle| x_i, \frac{y_i}{y_{Mi}}; \beta\right) + \varepsilon_i = X'_{it}\beta + v_i - u_i \quad (31)$$

The endogeneity of some regressors will lead to OLS being biased and inconsistent. This source of inconsistency can be dealt with by using GMM. However, the parameter estimates can still be inconsistent if u_i is heteroskedastic itself. To achieve consistent estimates, it is critical to ensure that chosen instruments do not include determinants of u_i. Suppose that we can find a vector of instruments M_i that satisfy the following moment condition:

$$E[M_i \cdot \varepsilon_i] = E[M_i \cdot (lny_{Mi} - X'_{it}\beta)] = E[m_i(\beta)] = 0 \quad (32)$$

The efficient two-step GMM estimator is then the parameter vector that solves:

$$\hat{\beta} = \arg\min[\textstyle\sum_{i=1}^{N} m_i(\beta)]' W^{-1} [\sum_{i=1}^{N} m_i(\beta)] \quad (32)$$

where W is an optimal weighting matrix obtained from a consistent preliminary GMM estimator. This optimal weighting matrix can take into account both heteroskedasticity and autocorrelation of the error term.[17]

In the *second* stage of the estimation procedure, distributional assumptions are invoked to obtain consistent estimates of the parameter(s) describing the standard deviations of allow v_i and u_i, conditional on the first-stage estimated parameters. Given that we are going to

[15] If there are more than two inputs, $\xi_i = (\xi_{2i}, \ldots, \xi_{Ji})$ follows a multivariate normal distribution.

[16] The endogeneity of some regressors will lead to least squares being biased and inconsistent. This source of inconsistency can be dealt with by using GMM. However, the parameter estimates can still be inconsistent if u_i is heteroskedastic itself. Indeed, a relevant issue that is often ignored when using OLS or GMM in a stochastic frontier framework is the endogeneity problem caused by the so-called "left-out variables" (Wang and Schmidt, 2002), which arises because variables influencing technical inefficiency are ignored when estimating the model. To achieve consistent estimates, it is critical to ensure that chosen instruments do not include determinants of u_i.

[17] If we allow v_i or u_i be heteroscedastic, an efficient GMM estimator is needed.

assume a particular distribution for the inefficiency term, both variances can be estimated using ML. The ML estimators are obtained by maximizing the likelihood function associated to the error term $\hat{\varepsilon}_i = lny_{Mi} - X'_{it}\beta$ that can be obtained from an estimate of the first-stage production equation (31). However, it should be pointed out that $\hat{\varepsilon}_i$ is a biased estimate of ε_i because $E(u_i) > 0$. We have two options to control for this bias. First, we can estimate the following (unrestricted) ML model:

$$\hat{\varepsilon}_i = \gamma_0 + v_i - u_i \tag{33}$$

where $\hat{\gamma}_0$ is an estimate of $E(u_i)$. If we assume that u_i follows a half-normal distribution, its mean value is equal to $\sqrt{2/\pi}\sigma_u$. Thus, the second option is estimating the above ML model with the following restriction $\gamma_0 = \sqrt{2/\pi}\sigma_u$.[18]

In the third stage we obtain the efficiency scores for each firm. From previous stages we have estimates of $\varepsilon_i = v_i - u_i$, which obviously contain information on u_i. The problem is to extract the information that ε_i contains on u_i. Jondrow *et al.* (1982) propose using the *conditional* distribution of the asymmetric random term u_i given the composed error term ε_i. The best predictor of u_i is the conditional expectation $E(u_i|\varepsilon_i)$ (see Kumbhakar and Lovell, 2000). Given our distributional assumptions, $E(u_i|\varepsilon_i)$ can be written as follows:

$$\hat{u}_i = E(u_i|\varepsilon_i) = \mu_* + \frac{\sigma_*\phi\left(\frac{\mu_{*i}}{\sigma_*}\right)}{\phi\left(\frac{\mu_{*i}}{\sigma_*}\right)} = E(u_i^*|\varepsilon_i) \tag{34}$$

where $\mu_* = -\varepsilon_i\sigma_u^2(\sigma_v^2 + \sigma_u^2)^{-1}$ and $\sigma_*^2 = \sigma_v^2\sigma_u^2(\sigma_v^2 + \sigma_u^2)^{-1}$. One might be tempted to validate the chosen specification of the inefficiency term by simply comparing the observed distribution of $\hat{u}_i$ to the assumed distribution for u_i. Wang and Schmidt (2009) show that this is not a good idea. To carry out this test we should compare the distribution of $\hat{u}_i$ and $E(u_i|\varepsilon_i)$. In this sense, they propose non-parametric Chi-square and Kolmogorov-Smirnov type statistics to perform this test properly. These authors also point out that, although $\hat{u}_i$ is the minimum mean squared error estimate of u_i, and it is unbiased in the unconditional sense $E(\hat{u}_i - u_i) = 0$, it is a shrinkage of u_i toward its mean. An implication of shrinkage is that on average we will overestimate u_i when it is small and underestimate u_i when it is large. This result, however, simply reflects the familiar principle that an optimal (conditional expectation) forecast is less variable than the term being forecasted.

Two comments are in order to conclude this section. First, it should be pointed out that, although we do not make any distributional assumptions on the noise and inefficiency terms when estimating the first-stage least squares equation, we still need the distributional assumptions to calculate the JLMS-type efficiency estimates based on Jondrow et al. (1982) formula, so that using OLS or GMM does not let researchers dispense with distributional assumptions altogether. Moreover, the computed efficiency scores rely in the end on the same

[18] The second-stage model can be also estimated by MM that relies on the second and third moments of the error term $\hat{\varepsilon}_i$ in equation (31). This approach takes advantage of the fact that, while the second moment provides information about both σ_v and σ_u, the third moment only provides information about the asymmetric random conduct term. Olson *et al.* (1980) showed using simulation exercises that the choice of the estimator (ML versus MM) depends on the relative values of the variance of both random terms and the sample size. When the sample size is large and the variance of the one-sided error component is small, compared to the variance of the noise term, ML outperforms MM. Second, the MM approach has some practical problems. Neglected heteroskedasticity in either or both of the two random terms causes estimates. Kumbhakar and Lovell (2000) pointed out that only the ML approach can be used to address this problem. Another practical problem arises when, in homoscedastic specifications of the model, the implied σ_u becomes sufficiently large to cause $\sigma_v < 0$, which violates the assumptions of the econometric theory.

distributional assumptions regardless of whether we use OLS or ML. Second, using a Hausman test of the difference between the ML and first-stage OLS equation to test distributional assumptions might not be a good idea. In principle, the ML estimator should be more efficient because it uses the distributional information, and the first-stage OLS estimator is likely to be consistent regardless of whether or not the inefficiency term follows a particular (homoscedastic) distribution. But, what about the ML estimator? This is not a trivial question. In particular, it is not clear whether the ML estimator is still consistent if we assume the wrong distribution for the inefficiency term. In the case that both estimators are consistent, we can use a Hausman test, but it will not necessarily show power if the ML is consistent too.

3.3. Distribution-free approaches

Firms' efficiency scores can also be computed without making specific distributional assumptions on the error components using the so-called *distribution-free approach*. In the following paragraphs, we present three methods that do not make distributional assumptions on either allow v_i or u_i .

3.3.1. COLS method

The Corrected Ordinary Least Squares (COLS) method was proposed by Winsten (1957) and can be used with cross-sectional or panel data sets. The estimation proceeds in two stages. In the first stage, we estimate the frontier parameters of (31) using OLS if the explanatory variables are exogeneous, or GMM if they are endogenous. At this stage, we obtain the zero-mean first-stage residuals as $\hat{\varepsilon}_i = lny_{Mi} - X'_{it}\hat{\beta}$. The value of $\hat{\varepsilon}_i$ can be greater, equal to, or less than zero. At the second stage, the estimated function is shifted upward to the extent that the function after the adjustment bounds all observations below. Once the residuals are adjusted upward, the frontier model becomes:

$$lny_{Mi} = \max(\hat{\varepsilon}_i) + X'_{it}\hat{\beta} - \hat{u}_i \tag{35}$$

and the inefficiency term is computed as:

$$\hat{u}_i = \max_i(\hat{\varepsilon}_i) - \hat{\varepsilon}_i \geq 0 \tag{36}$$

Notice that frontier model in (35) is deterministic in nature because any deviation from the frontier is now interpreted as inefficiency. This limitation can be addressed if panel data is available.

3.3.2. SS method

A fixed-effect estimator can be used to estimate the frontier model if panel data is available. In this case, it is possible to compute firm-specific efficiency scores without making specific distributional assumptions on the error components and using a stochastic or non-deterministic frontier framework.

Schmidt and Sickles (1984) assumed a production (distance) model with firm-specific intercepts that can be written as:

$$lny_{Mit} = \beta_0 + X'_{it}\beta + v_{it} - u_i = \alpha_i + X'_{it}\beta + v_{it} \tag{37}$$

where $\alpha_i = \beta_0 - u_i$ are firm-specific intercepts that are to be estimated along with the parameter vector β, and $X'_{it}\beta$ is the log of the frontier production (distance) function. Schmidt and Sickles (1984) showed that we can apply standard FE panel data estimation methods to estimate the firm-specific effects. Once $\hat{\alpha}_i$ are available, the following transformation is used to get time-invariant inefficiency scores for each firm:

$$\hat{u}_i = \max_i(\hat{\alpha}_i) - \hat{\alpha}_i \geq 0 \tag{38}$$

3.3.3. CSS method

To make the inefficiency term time-varying, Cornwell et al. (1990) suggest replacing α_i by $\alpha_{it} = \alpha_{0i} + \alpha_{1i}t + \alpha_{2i}t^2$. The model can be estimated using OLS if a set of firms' dummies and their interaction with t and t^2 are added to the model:

$$lny_{Mit} = \sum_{i=1}^{N}(\alpha_{0i} + \alpha_{1i}t + \alpha_{2i}t^2)D_i + X'_{it}\beta + v_{it} \quad (39)$$

Finally, $\hat{u}_{it}$ is obtained by:

$$\hat{u}_{it} = \hat{\alpha}_t - \hat{\alpha}_{it} = \max_i(\hat{\alpha}_{it}) - \hat{\alpha}_{it} \geq 0 \quad (40)$$

Notice that we can rewrite (39) using (40) as $lny_{Mit} = \hat{\alpha}_t + X'_{it}\beta + v_{it} + v_{it} - \hat{u}_{it}$. As $\hat{\alpha}_t$ changes over time, the CSS model allows implicitly for technical change, and the rate of technical change can be computed as $TC = \hat{\alpha}_t - \hat{\alpha}_{t-1}$.

4. More (advanced) topics and extensions

4.1. Observed environmental conditions

The concern about the inclusion of environmental variables (also called *contextual* or *z*-variables) has generated the development of several models either using parametric, nonparametric or semi-parametric techniques. Here we only mention the methods most frequently applied that include z-variables as determinants of firms' inefficiency.

The first methodological choice is whether we should incorporate the z-variables as either frontier determinants, determinants of firms' inefficiency, or as determinants of both the frontier and the inefficiency term. The key question that should be responded in order to include the z-variables as frontier determinants is whether a fully efficient firm will need to use more inputs to provide the same services or produce the same output level if an increase in a contextual variable represents a deterioration in the environment where it operates. In general, we should include as frontier drivers those variables that are fundamental to production. If they in addition make it more difficult or easier to manage the firm, they should be also treated as determinants of firms' inefficiency.

Like the two-stage DEA method, early papers aiming to understand firms' inefficiency using the SFA approach proceeded in two steps. In the first step, one estimates the stochastic frontier model and the firms' efficiency levels, ignoring the *z*-variables. In the second step, one tries to see how efficiency levels vary with *z*. It has long been recognized that such a two-step procedure will give biased results (see, for instance, Wang and Schmidt, 2002). The solution to this bias is a one-step procedure based on a heteroscedastic SFA model.

Once heteroscedastic SFA models are to be estimated, a second methodological choice appears: how to do it. Summaries of this literature can be found in Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014). The available options can be discussed using the general specification of the inefficiency term introduced by Álvarez et al. (2006): [19]

$$u_i \sim N^+(\mu(z_i), \sigma_u(z_i)) \quad (41)$$

where both the pre-truncation mean and standard deviation of the distribution might depend on the z-variables. According to this model, Álvarez et al. (2006) divide most heteroscedastic SFA models into three groups. In the *mean-oriented* models, it is assumed that the variance of the pre-truncated normal variable is homoscedastic and, thus, the contextual variables are

[19] The general models introduced by Wang (2002) and Lai and Huang (2010) are similar but they parameterize the pre-truncation mean of the distribution as a linear function of the *z*-variables.

introduced here through the pre-truncated mean. Following Battese and Coelli (1995), this specification can be written as:

$$u_i \sim N^+(\theta_0 + z_i'\theta, e^{\gamma_0}) \tag{42}$$

In contrast, in the *variance-oriented* models, it is assumed that the mean of the pre-truncated normal variable is homoscedastic and, hence, the environmental variables are treated as determinants of the variance of the pre-truncated normal variable. Following Caudill et al. (1995), this specification can be illustrated as:

$$u_i \sim N^+(0, e^{\gamma_0 + z_i'\gamma}) \tag{43}$$

In more *general* models, the contextual variables are introduced through both the mean and variance of the pre-truncated normal distributed random variable. Álvarez et al. (2006) and Lai and Huang (2010) proposed respectively exponential and lineal specifications for this model:

$$u_i \sim N^+\left(e^{\theta_0 + z_i'\theta}, e^{\gamma_0 + z_i'\gamma}\right) \tag{44}$$

$$u_i \sim N^+(\theta_0 + z_i'\theta, e^{\gamma_0 + z_i'\gamma}) \tag{45}$$

Some of the above models satisfy the so-called *scaling property* in the sense that the inefficiency term can be written as a deterministic (scaling) function of a set of efficiency covariates (h_i) times a one-sided random variable (u_i^*) that does not depend on any efficiency determinant. That is:

$$u_i = h(z_i'\gamma) \cdot u_i^* = h_i u_i^* \tag{46}$$

where e.g. u_i^* might follow a truncated normal or a half-normal distribution. For instance, the variance-oriented model in (43) has the scaling property due it can be rewritten as:

$$u_i = e^{z_i'\gamma} \cdot u_i^* \tag{47}$$

where $h_i = e^{z_i'\gamma}$ and $u_i^* \sim N^+(0, e^{\gamma_0})$. As Parmeter and Kumbhakar (2014) point out, the ability to reflect the scaling property requires that both the mean and the variance of the truncated normal are parameterized identically and with the same parameters in each parameterization. In this sense, the general model introduced by Álvarez et al (2006) also has the scaling property if we impose in (44) that $\theta = \gamma$. In this case, $h_i = e^{z_i'\gamma}$ and $u_i^* \sim N^+\left(e^{\theta_0}, e^{\gamma_0}\right)$.

The defining feature of models with the scaling property is that firms differ in their mean efficiencies, but not in the shape of the distribution of inefficiency. That is, the scaling property implies that changes in z_i affect the scale but not the shape of u_i. In this model u_i^* can be viewed as a measure of basic inefficiency which captures things like the managers' natural skills, which we view as random. How well these natural skills are exploited to manage the firm efficiently depends on other variables z_i, which might include the manager's education or experience, or measures of the environment in which the firm operates.

Although it is an empirical question whether or not the scaling property should be imposed, it has some features that make it attractive to some authors (see, e.g., Wang and Schmidt, 2002). Several authors have found the scaling property useful to remove individual fixed effects and still get a closed-form for the likelihood function (Wang and Ho, 2010), to address endogeneity issues (Griffiths and Hajargasht, 2016) or to relax the zero-rebound effect assumption in traditional demand frontier models (Orea et al., 2015).

As noted by Simar et al. (1994), Wang and Schmidt (2002), and Álvarez et al. (2006), the most fundamental benefit of the scaling property from a statistical point of view is that the stochastic frontier and the deterministic component of inefficiency can be recovered without requiring a specific distributional assumption on u_i^*. Indeed, if we take into account our specification of firms' inefficiency in (47) and define $\mu^* = E(u_i^*) \geq 0$, then taking expectations in (31) yields:

$$lny_{Mi} = X_i'\beta - h(z_i'\gamma) \cdot \mu^* + \varepsilon_i^* \tag{48}$$

where again and $X_i'\beta$ is the log of the frontier production (distance) function, and

$$\varepsilon_i^* = v_i - h(z_i'\gamma)[u_i^* - \mu^*] \tag{49}$$

The parameters in (49) can be estimated using nonlinear least squares as:[20]

$$(\hat{\beta}, \hat{\gamma}, \hat{\mu}^*) = \arg min \frac{1}{N} \sum_{i=1}^{N} [lny_{Mi} - X_i'\beta + h(z_i'\gamma)\mu^*]^2 \tag{50}$$

Given that ε_i^* is heteroscedastic, robust standard errors should be constructed to ensure valid inferences. The presence of μ^* in (50) implies that one cannot include a constant in h_i as this leads to identification issues (see Parmeter and Kumbhakar, 2014, p. 88). Interesting enough, μ^*cannot be estimated in a simple model where the inefficiency term is homoskedastic because, in this case, it cannot be distinguished from the intercept of the production frontier. As μ^* is multiplied by h_i in (50), we can get a separate estimate of both parameters.

In a second stage, distributional assumptions are invoked to obtain ML consistent estimates of the parameter(s) describing the variance and covariance of v_i, conditional on the first-stage estimated parameters. Notice that, if we assume that $u_i^* \sim N^+(0, \sigma_u)$, we have already got an estimate of σ_u using the first-stage estimate of μ^*as follows: $\hat{\sigma}_u = \hat{\mu}^* \sqrt{\pi/2}$. Thus, only σ_v should be estimated in the second-stage of the procedure. In a third stage, we can obtain the estimates of efficiency for each firm using the conditional expectation $E(u_i|\varepsilon_i^*)$.

All heteroscedastic frontier models above can be used to examine exogenous (marginal) effects on firm's expected inefficiency. These effects can be easily computed if the inefficiency term has the scaling property. For instance, assume u_i follows the heteroscedastic half-normal distribution in (43). In this case, the conditional expectation $E(u_i|z_i)$ is equal to $h_i \cdot E(u_i^*) = e^{z_i'\gamma} \cdot \left[\sqrt{2/\pi} e^{\gamma_0}\right]$. Thus, the marginal effect of z_i on $E(u_i|z_i)$ is:

$$\frac{\partial E(u_i|z_i)}{\partial z_i} = \gamma \cdot e^{z_i'\gamma} \left[\sqrt{2/\pi} e^{\gamma_0}\right] \tag{51}$$

In order to get non-monotonic effects, we could include quadratic terms, or estimate more general models with both heteroscedastic mean and variance. However, in the latter case, Wang (2002) shows that the marginal effects are complex functions of both γ and θ parameters.

4.2. Endogeneity issues

Endogeneity problems can arise in stochastic frontier models if the frontier determinants are correlated with the noise term, the inefficiency term or both. As noted by Kumbhakar et al. (2013), the endogeneity issue is typical in econometric models, especially when economic behaviours are believed to affect both inputs and/or outputs levels and inputs

[20] To impose $\mu^* \geq 0$ in practice, we could replace μ^* in (48) with e^{θ_μ}.

and/or outputs ratios.[21] Dealing with the endogeneity issue is relatively more complicated in a SFA framework than in standard regression models due to the special nature of the error term. Several authors have recently proposed alternative empirical strategies to account for endogenous regressors in SFA settings. In the next paragraphs we outline the main features of these methods, trying to identify their relative advantages and disadvantages.

Let us first assume that we are interested in estimating the following production model with endogenous regressors:

$$lny_{Mi} = X_i'\beta + v_i - u_i \tag{52}$$

$$X_i = z_i'\delta + \eta_i \tag{53}$$

where X_i is a vector of endogenous production drivers, and z_i is a vector of exogenous or instrumental variables. Equation in (53) can be viewed as a reduced form expression that links the endogenous variables with the set of instruments. The endogeneity problem arises if η_i is correlated with either v_i or u_i. In order to estimate consistently the frontier model (52), Guan et al. (2009) propose a two-step MM estimation strategy. In the first step, they suggest estimating the frontier parameters using a GMM estimator as long as valid instruments are found. In the second step, σ_v and σ_u are estimated using ML, conditional on the first-stage estimated parameters.

Instead of introducing instruments for these endogenous variables in an *ad hoc* fashion (e.g., temporal lags of inputs and outputs), Kumbhakar et al. (2013) and Malikov et al. (2015) bring additional equations for the endogenous variables from the first-order conditions of profitability (cost) maximization (minimization). They advocate using a system approach for two reasons. First, estimates of allocative inefficiencies can be obtained from the residuals of the first-order conditions. Second, since the first-order conditions contain the same technology parameters, their estimates are likely to be more precise (efficient). However, estimation of such a system requires availability of input and output prices. Their identification strategy also relies on competitively determined output and input prices as a source of exogenous variation.

Other authors make efforts to address the endogeneity problem in a fully ML estimation context. They use likelihood based instrumental variable estimation methods that rely on the joint distribution of the stochastic frontier and the associated reduced form equations in (53). The simultaneous specification of both types of equations has the advantage that it provides more efficient estimates of the frontier parameters as well as improvement in predicting the inefficiency term. For instance, Karakaplan and Kutlu (2013) assume that the error terms in (52) and (53) satisfy the following: [22]

$$\begin{pmatrix} \Omega_\eta^{-1/2}\eta_i \\ v_i \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} I_p & \rho\sigma_v \\ \rho'\sigma_v & \sigma_v^2 \end{bmatrix}\right) \tag{54}$$

where Ω_η is the variance-covariance matrix of η_i and ρ is a correlation vector between v_i and η_i. Based on (54), the equations in (52) and (53) can be written as:

$$lny_{Mi} = X_i'\beta + \tau(X_i - z_i'\delta) + \omega_i - u_i \tag{55}$$

[21] On the other hand, in cost (profit) settings, endogeneity problems might appear when the outputs' levels (prices) or input prices depend on random shocks and economic inefficiency. This might happen if firms are allocative inefficient, or firms have market power as, in this case, input/output prices are not set competitively in the market.

[22] In his model, the distribution of u_i is not allowed to have efficiency determinants.

where $\omega_i = (1 - \rho'\rho)v_i$ and $\tau = \sigma_v \rho' \Omega_\eta^{-1/2}$, which can be viewed as a correction term for bias. Note that $\omega_i - u_i$ is conditionally independent from the regressors given X_i and z_i. Hence, conditional on X_i and z_i, the distribution of the composed error term in (55) is exactly the same as their traditional counterparts from the stochastic frontier literature. They then show that the joint log-likelihood function of lny_{Mi} and X_i is given by:

$$lnLF = lnLF_{y|X} + lnLF_X \tag{56}$$

The first part of the log-likelihood function, $lnLF_{y|X}$, is almost the same as that of a traditional stochastic frontier model where the residual is adjusted by the $+\tau(X_i - z_i'\delta)$ factor. The second part, $lnLF_X$, is just the likelihood function of the reduced form equations in (53), that is the likelihood function of a multivariate normal variable. The likelihood function (56) can be maximized to obtain consistent estimates of all parameters of the model. However, if computational difficulties appear, one can use a two-step maximum likelihood estimation method. In the first stage, $lnLF_X$ is maximized with respect to Ω_η and δ. In the second stage, the rest of the parameters are estimated by maximizing $lnLF_{y|X}$ taking the estimates of Ω_η and δ as given.[23]

The abovementioned ML model does not address the potential correlation with the inefficiency term, and neither does it assure consistency of parameter estimates when η_i is correlated with both v_i and u_i. Amsler et al. (2016) is the first paper to allow endogeneity of the inputs with respect to statistical noise and inefficiency separately. They propose using a (Gaussian) copula in order to specify the joint distribution of these three random variables.[24] One obvious difficulty with this approach is the need to specify a copula. Another difficulty of this approach is that it may be computationally challenging. Tran and Tsionas (2015) also use a Gaussian copula function to directly model the dependency of the endogenous regressors and the composed error without using instrumental variables. Consistent estimates can be obtained by maximizing the likelihood function in a two-step procedure. The first step requires, however, using numerical integration as in Amsler et al. (2016).[25]

4.3. Unobserved heterogeneity

Many industries worldwide are incentive regulated. The aim is to provide firms with incentives to improve their efficiency and to ensure that consumers benefit from the gains. As regulators reward or penalise firms in line with their respective (in)efficiency levels, the reliability of these scores is crucial for the fairness and effectiveness of the regulatory framework. Obtaining reliable measures of firms' inefficiency requires controlling for the different environmental conditions under which each firm operates. This is particularly important in the case of benchmarking of electricity, gas, and water networks where the results of efficiency analysis have important financial implications for the firms. However, there are many characteristics (e.g., geography, climate or network characteristics) that affect firms' production (costs) but which are *unobserved* or *omitted* variables.

---

[23] However, the standard errors from this two-stage method are inconsistent because the estimates are conditional on the estimated error terms from the first stage. Kutlu (2010) suggests using a bootstrapping procedure in order to get the correct standard errors.

[24] A copula is a multivariate probability distribution for which the marginal probability distribution of each variable is uniform.

[25] In the abovementioned papers, there were no environmental variables determining firms' inefficiency. Amsler et al (2017) provides a systematic treatment of endogeneity in stochastic frontier models and allows environmental variables to be endogenous because they are correlated with either the statistical noise or the basic inefficiency term or both.

Several statistical methods have been developed in the SFA literature to address this issue. A simple or naïve strategy is the sample separation approach. Estimation of the technology is carried out in two stages. First, the sample observations are classified into several groups. In the second stage, separate analyses are carried out for each class, conditional on the first-stage (maybe ad-hoc) sample separation. More sophisticated and popular approaches to deal with omitted variables use panel data, random coefficients, latent class models, or spatial econometrics.

4.3.1. Panel data models

For instance, the True Fixed/Random Effects models introduced by Greene (2005) capture the unobserved heterogeneity through a set of firm-specific intercepts α_i:

$$lny_{Mit} = \alpha_i + X'_{it}\beta + v_{it} - u_{it} \tag{57}$$

If we treat α_i as fixed parameters which are not part of inefficiency, then the above model becomes the "True Fixed Effects" (TFE) panel stochastic frontier model. The model is labelled as "True Random Effects" model when α_i is treated as a time-invariant random variable. Estimation of the model in (57) is not easy. When the number of firms is too large, the model encounters the incidental parameter problems. This problem appears when the number of parameters to be estimated increases with the number of cross-sectional observations in the data. In this situation, consistency of the parameter estimates is not guaranteed even if $N \to \infty$.

Wang and Ho (2010) solve the problem in Greene (2005) using temporal transformations of (57). In order to remove time-invariant firm-specific effects , they carried out first-differences and within transformations of the model. Their within transformation of the inefficiency term (see eq. 24) is reproduced below with our notation:

$$u_{it}^{w} = u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it} = u_{it} - u_{i.} \tag{58}$$

As u_{it}^{w} is the difference of "two" one-sided error terms, the distribution of u_{it}^{w} is not known if u_{it} is independently distributed over time. To get a closed form for the likelihood function, they assumed that the inefficiency term u_{it} possesses the scaling property so that it can be multiplicatively decomposed into two components as follows:

$$u_{it} = h(z_{it}, \delta) \cdot u_i^{*} = h_{it} \cdot u_i^{*} \tag{59}$$

where $h_{it} \geq 0$ is a function of firm exogenous variables, and $u_i^{*} \geq 0$ is a firm-specific and *time-invariant* inefficiency term which captures aspects such as the manager's natural skills which are viewed as random. Note that this implies that the within-transformed inefficiency term in (58) can be rewritten as:

$$u_{it}^{w} = \left(h_{it} - \frac{1}{T}\sum_{t=1}^{T} h_{it}\right) \cdot u_i^{*} \tag{60}$$

Note that the distribution of u_i^{*} is not affected by the within-transformation. This key aspect of their model enabled them to get a tractable likelihood function for their transformed model.

Note that the TFE model and WH models capture the unobserved heterogeneity through a set of firm-specific intercepts α_i. If we use the SS method, the adjusted individual effects provide a measure of persistent (time-invariant) inefficiency. In order to separate persistent inefficiency from both time-invariant unobserved heterogeneity and transient (time-varying) inefficiency, Greene propose estimating a model with four error terms:

$$lny_{Mit} = \beta_0 + X'_{it}\beta + \alpha_i + v_{it} - (u_{\mathrm{i}} + \tau_{it}) \tag{61}$$

where α_i captures time-invariant unobserved heterogeneity, v_{it} is the traditional noise term that follows a normal distribution, u_{i} is one-sided error term capturing persistent inefficiency, and τ_{it} is one-sided error term capturing transient inefficiency. Estimation of the model in (61) can be undertaken in a single stage ML method based on the distributional assumptions on the four error terms (Colombi et al., 2011).[26] Kumbhakar et al. (2015) consider a simpler multi-stage procedure on the model is rewritten as:

$$lny_{Mit} = \beta_0^* + X_{it}'\beta + \alpha_i^* + \omega_{it} \tag{62}$$

where $\beta_0^* = \beta_0 - E(u_{\mathrm{i}}) - E(\tau_{it})$, $\alpha_i^* = \alpha_i - u_{\mathrm{i}} + E(u_{\mathrm{i}})$ and $\omega_{it} = v_{it} - \tau_{it} + E(\tau_{it})$. This model can be easily estimated in three stages. In the first stage, we estimate (62) using a FE or RE estimator and get the first-stage fixed-effects ($\hat{\alpha}_i^*$) and residuals ($\hat{\omega}_{it}$). In a second stage, we estimate a standard SFA model regressing $\hat{\omega}_{it}$ on an intercept, which can be interpreted as an estimate of $E(\tau_{it})$. Using the Jondrow et al. (1082) formula, we decompose $\left(\hat{\omega}_{it} - \hat{E}(\tau_{it})\right)$ into $\hat{v}_{it}$ and $\hat{\tau}_{it}$. In the third stage, we estimate a SF model regressing $\hat{\alpha}_i^*$ on an intercept, which can be interpreted as an estimate of $E(u_{\mathrm{i}})$. Using again Jondrow et al. (1982), we next decompose $\left(\alpha_i^* - \hat{E}(u_{\mathrm{i}})\right)$ into $\hat{\alpha}_i$ and $\hat{u}_i$.

To conclude this subsection, it is worth mentioning that the above panel data models only use the temporal (i.e. within) variation contained in the data to estimate the coefficients of the main production drivers. This is quite problematic in many applications because many important determinants of firm costs (production) are persistent or slow changing variables (such as the energy delivered or number of customers in electricity distribution).

4.3.2. Latent class models

Possible differences among firms associated with their use of different technologies are also often addressed using latent class models. The latent class stochastic frontier (hereafter LCSF) models combine the stochastic frontier approach with a latent class structure (see Orea and Kumbhakar, 2004; Greene, 2005, for some applications). A conventional LCSF model assumes there is a finite number of technologies (classes) underlying the data and allocates probabilistically each firm in the sample to a particular technology.[27]

Let us first assume that there are *J* different technologies, and that each firm belongs to one, and only one, of these technologies. Conditional on technology *j* (=1,…,*J*), the general specification of the LCSF model can be written as follows:

$$lny_{Mi} = X_i'\beta_j + v_{i|j} + u_{i|j} \tag{63}$$

where $v_{i|j} \sim N(0, \sigma_{vj})$ is a noise term that follows a normal distribution, and $u_{i|j} \sim N^+(0, \sigma_{uj})$ is a one-sided error term capturing firms' inefficiency.[28] Given that the researcher lacks knowledge as to whether a particular firm belongs to class *j* or another, the class-membership probability should be estimated simultaneously alongside other parameters of the model.

[26] Greene and Fillipini (2016) proposed a simulation-based procedure to circumvents many of the challenges that appears when estimating the model by brute force maximization.

[27] The LCSF model is similar to the stochastic frontier model with random coefficients introduced by Tsionas (2002), in the sense that a latent class model can be viewed as a discrete approximation to a (continuous) random coefficient model (see Greene, 2005, p. 287).

[28] Orea and Jamasb (2017) assumed the existence of two behavioral classes: fully efficiency and inefficient. While in the "inefficient" class it is assumed that $u_{i|j}$ follows a half-normal distribution, the "fully efficient" class is defined by imposing that the variance of the pre-truncated normal distribution is zero, i.e. $\sigma_{uj}^2 = 0$.

Following Greene (2005), the class probabilities are parameterized as a multinomial logit function:

$$\Pi_{ij}(\gamma_j) = \frac{exp(q_i'\gamma_j)}{1+exp(q_i'\gamma_j)}, \qquad j = 1, \dots, J-1 \tag{64}$$

where q_i is a vector of firm-specific variables. The last probability is obtained residually taking into account that the sum of all probabilities should be equal to one. The *unconditional* likelihood for firm i is obtained as the weighted sum of their technology-specific likelihood functions, where the weights are probabilities of technology-class membership, Π_{ij}. That is:

$$LF_i(\theta) = \sum_{j=1}^{J} LF_{i|j}(\beta_j, \sigma_{vj}, \sigma_{uj})\Pi_{ij}(\gamma_j) \tag{65}$$

where θ encompasses all parameters. The overall likelihood function can be written as:

$$ln\, LF\ (\theta) = \sum_{i=1}^{N} ln\, LF_i\ (\theta) = \sum_{i=1}^{N} ln\{\sum_{j=1}^{J} LF_{i|j}(\beta_j, \sigma_{vj}, \sigma_{uj})\Pi_{ij}(\gamma_j)\} \tag{66}$$

Maximizing the above maximum likelihood function gives asymptotically efficient estimates of all parameters. The estimated parameters can then be used to compute (unconditional) posterior class membership probabilities for each technology. The posterior probabilities can be used to allocate each firm to a technology-class with highest probability.

4.3.3. Spatial frontier models

A common feature of the above approaches is that they ignore the spatial structure of the data. Orea et al. (2018) advocate using a different empirical strategy to account for the unobserved differences in environmental conditions based on firms' geographic location. Indeed, as many unobservable variables are likely to be spatially correlated (such as weather and geographic conditions, population structure, electricity demand patterns, input prices, etc), an alternative empirical strategy emerges. Their spatial-based approach can be used in panel data settings. Indeed, as they utilise different (spatial vs. temporal) dimensions of our data, they can be viewed as *complementary* approaches to deal with unobserved variables.

Orea et al. (2018) proposed a frontier model with cross-sectional correlation in the noise term, which can be written assuming a single input as:

$$lny_{Mi} = X_i'\beta + v_i - u_i = X_i'\beta + (z_i + \omega_i) - u_i \tag{67}$$

$$z_i = \lambda W_i z \tag{68}$$

where z_i represent *unobserved* environmental variables that are spatially correlated, and ω_i is the traditional non-spatially correlated noise term, z is a vector of $Nx1$ unobserved environmental variables, W_i is a known $1xN$ spatial weight vector with elements that are equal to zero if a particular firm j is not a neighbour of firm i and equal to one if the two firms are neighbours – i.e. the service areas of the firms are adjacent. The term λ is a coefficient that measures the degree of spatial correlation between the unobserved environmental variables. Hence the spatial effects estimated in this model lack an economic interpretation as they are completely "spurious". Equation (67) can be alternatively rewritten as follows:

$$lny_{Mi} = X_i'\beta + \lambda W_i lny_M + W_i X'(-\lambda\beta) + \tilde{\omega}_i - \tilde{u}_i \tag{69}$$

where $\tilde{\omega}_i = \omega_i - \lambda W_i \omega$, $\tilde{u}_i = u_i - \lambda W_i u$, lny_M is a vector of $Nx1$ production levels, X is a vector of $Nx1$ explanatory variables, u is $Nx1$ vectors of the firms' inefficiency terms, and ω is again $Nx1$ vectors of the firms' non-spatially correlated noise terms.

Several comments are in order with respect to this specification. First, in contrast to (67), equation (69) is a model that now includes a set of spatially lagged variables, i.e. $W_i lny_M$

and W_iX. Therefore, equation (69) resembles a conventional spatial econometric model. However, in (67), only one additional coefficient is estimated, and the coefficient of the spatially lagged dependent variable should not be interpreted as the effect of neighbours' production on the production of a particular firm. Rather, λ is measuring the spatial correlation between the unobserved or omitted variables in our sample. On the other hand, it is worth mentioning that (69) is similar to the Durbin Stochastic Frontier (SDF) model introduced recently by Glass et al. (2016) in which they propose estimating the following model:

$$lny_{Mi} = X_i'\beta + \lambda W_i lny_M + W_i X'\theta + v_i - u_i \tag{70}$$

It is easily observable that our spatial model in (69) and the SDF model in (70) differ in two important aspects. First, the set of parameters θ in the SDF model is not restricted to be equal to $-\lambda\beta$. In this sense, our spatial model in (5) is nested in the SDF model. However, no spatially correlated omitted (random) variables are explicitly modelled in the SDF model. While the spatial spillovers in Glass et al. (2016) have an economic or causal interpretation, the spatial spillovers in our spatial model are simply associated with the omitted variables.

Orea et al. (2018) discuss how to estimate (69) taking into account that this model includes two spatially correlated error terms. They propose a two-step procedure. In the first step, equation (69) is estimated ignoring the spatial and frontier structure of the composed error term. The degree of spatial correlation of omitted variables (i.e. parameter λ) and other coefficients of the frontier model are estimated using GMM because the spatially lagged dependent variable is endogenous. The estimated λ parameter is then used to get a predicted value for z_i. In the second step, they estimate (67) once the original omitted variable z_i is replaced with its predicted counterpart.

Orea and Álvarez (2017) develop a cross-sectional (spatial) frontier model that explicitly allows for cross-sectional (spatial) correlation in both noise and inefficiency terms**.** Their model can be written as:

$$lny_{Mi} = X_i'\beta + \tilde{v}_i(\rho) + \tilde{u}_i(\tau) \tag{71}$$

where both error terms are cross-sectionally correlated using spatial moving average (SMA) or spatial autoregressive (SAR) spatial stochastic processes. The coefficients ρ and τ measure the degrees of cross-sectional (spatial) correlation between firms' noise and inefficiency terms respectively.[29] In a SMA specification of the model, the noise and inefficiency terms are defined as $\tilde{v}_i = v_i + \rho W_i v$, and $\tilde{u}_i = u_i + \tau W_i u$. A SAR specification for the two error terms can be expressed as: $\tilde{v}_i = v_i + \rho W_i \tilde{v}$, and $\tilde{u}_i = u_i + \tau W_i \tilde{u}$.

Note that (71) has the structure of a traditional SFA model as it includes a noise term ($\tilde{v}_i$) and an inefficiency term ($\tilde{u}_i$). However, the above model cannot be estimated using full maximum likelihood because the distribution of $\tilde{u}_i$ is *generally* not known *if* we assume that u_i is independently distributed across firms (see, for instance, Wang, 2003). To address this issue, Areal et al. (2012), Tsionas and Michaelides (2016), and Schmidt et al. (2009) proposed several computational algorithms based on Gibbs sampling or simulated ML. In contrast, Orea and Álvarez (2017) assumed that the basic inefficiency term u_i possesses the scaling property, but we replace Wang and Ho (2010)'s firm-specific term u_i^* with an industry-specific term u^*:

$$u_i = h(z_i, \delta) \cdot u^* = h_i \cdot u^* \tag{72}$$

[29] In a SMA specification of the model, the noise and inefficiency terms are defined as $\tilde{v}_i = v_t + \rho W_i v$, and $\tilde{u}_i = u_i + \tau W_i u$. A SAR specification for the two error terms can be expressed as: $\tilde{v}_i = v_i + \rho W_i \tilde{v}$, and $\tilde{u}_i = u_i + \tau W_i \tilde{u}$.

where $h_i \geq 0$ is again function of firm exogenous variables, and $u^* \geq 0$ is an industry-specific inefficiency term. For simplicity, Orea and Álvarez (2017) assume that $u^* \sim N^+(0, \sigma_u)$.[30] The above specification of u_i implies that the SMA-transformed inefficiency term can be written as:

$$\tilde{u}_i = u_i + \tau W_i u = \left(h_i + \tau \frac{1}{n_i} \Sigma_{j \in A_i} h_j\right) u^* = \tilde{h}_i \cdot u^* \tag{73}$$

or, in simpler notation:

$$\tilde{u} = (I_N + \tau W)u = M_\tau u = M_\tau h u^* = \tilde{h} \cdot u^* \tag{74}$$

where $h = (h_1, \ldots, h_N)$, and $\tilde{h} = (\tilde{h}_1, \ldots, \tilde{h}_N)$ are Nx1 vectors of *idiosyncratic* and *generalized* scaling functions, respectively. If the inefficiency term instead follows a SAR process, we just need to replace $M_\tau = I_N + \tau W$ with $M_\tau = (I_N - \tau W)^{-1}$. Regardless of whether SMA or SAR processes are assumed, the half-normal distribution of u^* is not affected by the cross-sectional transformation. This is the crucial aspect of the model that enables Orea and Álvarez (2017) to get a tractable likelihood function that can be maximized using standard software. In this sense, the proposed model can be viewed as a new application of the scaling property in SFA analyses. Moreover, some portions of the model can also be estimated using non-linear least squares (NLLS).

4.4. Dynamic efficiency

The empirical literature on efficiency was initially developed under a static theory of the firm. However, the decision-making process followed by producers is quite often dynamic in nature. Rigidities derived from the nature of some inputs, regulations, transaction costs, information failures and other adjustment costs may prevent firms from moving instantly towards long-run optimal conditions. When these constraints are taken into consideration, it could very well result that being on the production and cost frontier constantly may not be the optimal long-run strategy. Moreover, in this context, firms may not only find it optimal to remain inefficient in the short-run, but also their inefficiency may persist from one period to the next. Two different approaches have been used in the literature to incorporate the dynamic nature of the decision-making process into efficiency analyses: reduced-form models and structural models.[31]

4.4.1. Reduced-form models

The *reduced-form models* do not define explicitly a mathematical representation of dynamic behaviour of the firm but recognize a persistence effect of firms' inefficiency over time and specify its evolution as an autoregressive process. For instance, Tsionas (2006) departs from a typical stochastic production frontier of the following form:

$$lny_{Mit} = X'_{it}\beta + v_{it} + lnET_{it} \tag{75}$$

where $ET_{it} = e^{-u_{it}} \leq 1$ is the usual technical efficiency of firm *i* in period *t*. To avoid the complications inherent in the specification of autoregressive processes on non-negative

[30] As the random inefficiency component in (72) does not vary across firms, consistency of σ_u can be obtained if we use a panel data set and $T \to \infty$.

[31] For a more comprehensive review of this literature see Emvalomatis (2009).

variables, Tsionas (2006) converts the technical efficiency term into an autoregressive form using $s_{it} = \ln(-lnET_{it})$ instead of directly $lnET_{it}$:[32]

$$s_{it} = z'_{it}\delta + \rho s_{it-1} + \xi_{it} \tag{76}$$

The distinguishing feature of (58) is that past values of efficiency determine the value of ET_{it}. Estimating the above dynamic stochastic frontier model is far from simple. While Tsionas (2006) estimate the model using Bayesian techniques, Emvalomatis et al. (2011) use Kalman filtering techniques and proceed to estimation by maximum likelihood.

4.4.2. Structural models

The *structural models* that make explicit assumptions regarding the objective of the firm. For instance, the objective of the firm is often assumed to be the maximization of the following intertemporal problem (see Tovar and Wall, 2016):

$$\begin{aligned} W(y,K,x,w,c) &= \min_{I,x} E^t \left\{ \int_0^\infty e^{-rt}(w'x + cK)dt \right\} \\ s.t. \qquad \dot{K} &= I - \delta K \\ \vec{D}(y,K,x,I,-g_x,g_I) &\geq 0 \end{aligned} \tag{77}$$

where w is the vector of variable input prices, c is the capital rental price, I is gross investment, and r is the discount rate. In this formulation, the objective of the firm is to minimize the present value of costs. The choice variables are the levels of variable inputs (x) to be employed and the level of investment in quasi-fixed inputs (I). While the first restriction describes the evolution of capital through time, the second restriction is a *dynamic* representation of technology in terms of a directional distance function. Given the level of quasi-fixed inputs, this function describes the vectors of outputs that can be produced from a given vector of variable inputs and gross investment. Depending upon the orientation of the distance function, adjustment costs are implicit in higher variable inputs or lower output. Regardless of the dynamic specification, it should be noted that they all indicate that, in the presence of adjustment costs in quasi-fixed inputs, static measures do not correctly reflect inefficiency.

Serra et al. (2011) and Tovar and Wall (2016) used the adjustment cost framework of Silva and Lansink (2013), but instead of DEA they carried out a parametric estimation generalizing the static input-oriented directional distance function introduced by Färe et al. (2005). They all use a quadratic functional form for the directional distance function because it is easy to impose the above translation property. Their dynamic directional distance function is also input oriented. Therefore, $\vec{D}(y,K,x,I,-g_x,g_I)$ in (77) represents the maximum contraction of variable inputs and the maximum expansion of gross investments that keeps the combination of variable inputs and gross investments inside the input requirement set. Setting $(g_x, g_I) = (1,1)$, their dynamic input-oriented directional distance function can be written as:

$$\vec{D}(y_{it},K_{it},x_{it}-\alpha_{it},I_{it}+\alpha_{it}) = \vec{D}(y_{it},K_{it},x_{it},I_{it}) - \alpha_{it} \tag{78}$$

This simply states that if investment is expanded by α_{it} and input contracted $-\alpha_{it}$, the value of the distance function will be reduced by α_{it}. The above papers set $\alpha_{it} = I_{it}$. Stochastic estimation is accomplished by maximum likelihood procedures in Tovar and Wall (2016). Estimating the above directional distance function only provides estimates of technical

[32] Alternatively, Emvalomatis et al. (2011) define $s_{it} = \ln(ET_{it}/(1-ET_{it}))$ as the latent-state variable. In this specification, ρ measures the percentage change in the efficiency to inefficiency ratio that is carried from one period to the next.

inefficiency. To get cost efficiency scores in a dynamic framework, they propose estimating the following (quadratic) cost frontier model:

$$C_{it} = rW(\cdot) - W_K(\cdot)\dot{K}_{it} + v_{it} + u_{it} \tag{79}$$

where C_{it} is observed cost (normalized by a variable input price), $W(y_{it}, K_{it}, w_{it})$ is optimum cost, $W_K(\cdot)$ is its derivative with respect to the capital stock; v_{it} is white noise and u_{it} is a one-sided term measuring firms' cost inefficiency. The dynamic directional distance function (78) allows estimating technical inefficiency of both variable and quasi-fixed inputs. The parametric dynamic cost model (79) allows estimating the dynamic cost inefficiency defined as the difference between the observed shadow cost of input use and the minimum shadow cost. Finally, an allocative inefficiency score can be obtained as the difference between dynamic cost inefficiency and dynamic technical inefficiency.

4.5. Production risk.

Most of the literature measuring firms' production performance lacks an explicit recognition that production takes place under conditions of uncertainty. Although SFA models are stochastic, their stochastic elements arise primarily from econometric concerns (measurement error, missing variables) and not as an endogenous response to the stochastic environment in which firms operate. Ignoring uncertainty in efficiency and productivity analyses may have remarkable welfare and policy implications, which serve to jeopardize our interpretation of the efficiency measures and also bias our representation of the stochastic technology. This may be a serious issue in many applications, such as agriculture, fishing or banking where production uncertainty is relatively high.

Several approaches have been proposed in the applied literature to take these factors into account and thereby give a fuller picture of firms' performance under production/demand uncertainty. For many years the standard tool for analysing firms' performance under production risk has been the simple production function with heteroskedastic error terms representing risk (e.g., Just and Pope, 1978). Kumbhakar (2002), among others, extended this framework and proposed estimating the following single-output SFA model:

$$y_i = f(x_i, \beta) + g(x_i, \lambda)\{v_i - u_i\} \tag{80}$$

where $g(x_i, \lambda)$ is the output risk function. If the variance of the composed random term is normalized to 1, the variance of output is therefore $g(x_i, \lambda)$. In this framework, an input is risk-increasing (reducing) (neutral) according to $\partial g(x_i, \lambda)/\partial x_i > (<)(=)0$. Kumbhakar assumed later on that producers maximize the expected utility of anticipated profits. Assuming a single input, the first-order condition of the above problem can be expressed as:

$$\frac{\partial f(x_i,\beta)}{\partial x_i} = w - \theta(\cdot)\frac{\partial g(x_i,\beta)}{\partial x_i} \tag{81}$$

where w is the input price relative to the output price, and $\theta(\cdot)$ is a risk preference function that measures firms risk aversion.[33] This function takes values less than, equal to or higher than zero when producers are risk-averse, risk-neutral or risk-loving, respectively. Risk aversion coefficients can be estimated from this equation (or a system of equations in the case of more inputs) once the mean and variance marginal products are replaced by their predicted values from the prior SFA model. The distinctive feature of this type of model is the difficulty in

[33] The coefficient of risk aversion in this equation can be viewed as a measure of overall risk preferences regarding both noise and inefficiency terms.

deriving an algebraic form of the risk preference function that keeps the model simple for estimation purposes.[34]

A common feature of the previous model is that they it is developed using standard stochastic frontier models that are too simple to account properly for the stochastic elements of the producer decision environment. In this sense, O'Donnell et al. (2010) show that the application of standard methods of efficiency analysis to data arising from production under uncertainty may give rise to spurious findings of efficiency differences between firms. To deal with this issue, Chambers and Quiggin (2000) found it convenient to treat uncertainty as a discrete random variable and proposed to model uncertainty in terms of a state-contingent technology, where each state represents a particular uncertain event. The state-contingent approach recognizes that actions (input choices) can have different consequences with different states of nature, whereas the role that inputs play remains the same regardless of which state occurs in standard stochastic production models.

Empirical application of the state contingent approach has proved difficult because most of the data needed to estimate these models are lost in unrealized states of nature (i.e., outputs are typically observed only under one of the many possible states of nature). O'Donnell and Griffiths (2006) show how to estimate state-contingent models using a latent class model approach if the technology is "output-cubical" in the terminology of Chambers and Quiggin (2000).[35] In this case, the production technology can be described by the set of state-contingent production functions:

$$lny_i = \alpha_s + f(x_i, \beta) + v_{is} - u_{is} \tag{82}$$

where α_s is a state-varying intercept that allows expected log-output to vary across the states of nature. The standard deviation of v_{is} is assumed state-dependent. Technically inefficiency will also be expected to differ across states. The above model can be viewed as a conventional stochastic frontier model with state-specific parameters where the underlying (latent) state of nature that has produced each observation is not observed. For this reason, the above authors nest the above model into a latent class model (LCM) structure, where both state-specific production functions and the probabilities for the realization of each state are estimated simultaneously by ML techniques.[36]

4.6. Total factor productivity decomposition.

An estimated distance function can constitute the building block for the measurement of productivity change and its decomposition into its basic sources. First, let us add a *t* superscript to all variables of the output distance function (19) and a time trend to capture

[34] Orea and Wall (2012) used the above framework in order to show that increases in productivity, measured by a ratio of output to inputs, and welfare changes do not necessarily follow the same path when we recognize that production takes place under conditions of uncertainty and firms are not risk-neutral.

[35] Chavas (2008) proposes a method that allows the researcher to test whether or not the state-contingent technology is 'output-cubical' if the states are independently distributed across observations. The main limitation of this method is that it focuses exclusively on the observed outputs. As such, the approach neglects the potential outputs that could have been obtained had nature selected different states.

[36] Note that the elasticity of expected output with respect to the input in (63) is state-invariant. This property may be implausible in some production contexts (e.g., irrigation in rainy and dry seasons). If we allow the slope coefficients in (63) to vary across states of nature, an identification (or labelling) problem arises. If there are only two different states of nature, which class should be labelled as a 'bad' or 'good' state? To solve the identification problem, O'Donnell and Griffiths (2006) suggest scaling the inputs so that $x_i = 0$ at the sample mean. The state with the lowest (highest) α_s will be labelled as 'bad' ('good') state. In this case, however, this labelling is local in the sense that it is only valid for the 'representative' firm. O'Donnell and Griffith (2006) rely on Bayesian estimation to address the identification problem and impose the labelling restriction globally.

technological changes over time. Taking into account that $u_{it} = -lnD_{it}$, the distance function in period *t* can be rewritten as:

$$lnD_{it} = lny_{Mit} + \ln D\left(x_{it}, \frac{y_{it}}{y_{Mit}}, t, \beta\right) + v_{it} = \ln D(x_{it}, y_{it}, t, \beta) + v_{it} \quad (83)$$

If we take first differences, we get:

$$\Delta lnD_{it} = \Delta \ln D(x_{it}, y_{it}, t, \beta) + \Delta v_{it} \quad (84)$$

As the average change in the noise term tend to vanish over time, we hereafter ignore Δv_{it} for notational ease. We next assume that the above distance function has a Translog form. Since the Translog distance function is quadratic in logs, the change in the value of the distance function can be decomposed as:

$$\Delta \ln D(x_{it}, y_{it}, t, \beta) = \frac{1}{2}\sum_{m=1}^{M}\left(\varepsilon_{mi}(t) + \varepsilon_{mi}(t-1)\right)\Delta lny_{mit}$$
$$+\frac{1}{2}\sum_{j=1}^{J}\left(\varepsilon_{ji}(t) + \varepsilon_{ji}(t-1)\right)\Delta lnx_{jit} + \frac{1}{2}\left(\varepsilon_t(t) + \varepsilon_t(t-1)\right) \quad (85)$$

where $D(t)$ is short for $D(x_{it}, y_{it}, t, \beta)$, $\varepsilon_{mi}(t) = \frac{\partial lnD(t)}{\partial lny_{mi}}$ is the elasticity of the distance function with respect to y_{mi}, $\varepsilon_{ji}(t) = \frac{\partial lnD(t)}{\partial lnx_{ji}}$ is the elasticity of the distance function with respect to x_{ji}, and $\varepsilon_t(t) = \frac{\partial lnD(t)}{\partial t}$ is the rate of technical change evaluated at period t. In order to measure total factor productivity changes, Orea (2002) proposed the following Generalized Malmquist Productivity Index:

$$lnG_{t,t-1} = \frac{1}{2}\sum_{m=1}^{M}\left(\frac{\varepsilon_{mi}(t)}{\sum_{m=1}^{M}\varepsilon_{mi}(t)} + \frac{\varepsilon_{mi}(t-1)}{\sum_{m=1}^{M}\varepsilon_{mi}(t-1)}\right)\cdot \Delta lny_{mit}$$
$$-\frac{1}{2}\sum_{j=1}^{J}\left(\frac{\varepsilon_{ji}(t)}{\sum_{j=1}^{J}\varepsilon_{ji}(t)} + \frac{\varepsilon_{ji}(t-1)}{\sum_{j=1}^{J}\varepsilon_{ji}(t-1)}\right)\cdot \Delta lnx_{jit} \quad (86)$$

Notice that in (86) we have not imposed the linear homogeneity in outputs of the distance function $\sum_{m=1}^{M}\varepsilon_{mi}(t) = 1$ in order to show that it can also be used with input distance functions. Inserting (66) into (65), Orea (2002) obtained the following parametric decomposition of the Malmquist productivity index (67): [37]

$$lnG_{t,t-1} = \Delta lnD_{it} - \frac{1}{2}\left(\varepsilon_t(t) + \varepsilon_t(t-1)\right)$$
$$+\frac{1}{2}\sum_{j=1}^{J}\left(\frac{\varepsilon_{ji}(t)}{\sum_{j=1}^{J}\varepsilon_{ji}(t)}EE(t) + \frac{\varepsilon_{ji}(t-1)}{\sum_{j=1}^{J}\varepsilon_{ji}(t-1)}EE(t-1)\right)\cdot \Delta lnx_{jit} \quad (87)$$

where $EE(t) = -\sum_{j=1}^{J}\varepsilon_{ji}(t) - 1$ is a measure of firms' economies of scale. Equation (87) provides a meaningful decomposition of a total factor productivity indicator into changes in technical efficiency (TE), technical change (TC) and a scale effect (SE). The first term measures changes in technical efficiency over time. The negative sign of the second term transforms technical progress (regress) into a positive (negative) value. The scale term relies on scale elasticity values and on changes in input quantities, and therefore it vanishes under the assumption of constant returns to scale or constant input quantities.

[37] A similar decomposition can be obtained from a parametric directional distance function using a Luenberger productivity index (see Färe et al., 2008; p. 593).

It should be pointed out that the above decomposition does not individualize any output or input mix effect. However, an input mix effect can be easily obtained if we measure the scale effect with respect to the *average* input change, instead of the change of each input. In this case, the scale effect in (87) can be in turn decomposed in a *pure* scale effect and a term measuring relative changes in the input mix:

$$SE = \left\{ \frac{1}{2} \sum_{j=1}^{J} \left(\frac{\varepsilon_{ji}(t)}{\sum_{j=1}^{J} \varepsilon_{ji}(t)} EE(t) + \frac{\varepsilon_{ji}(t-1)}{\sum_{j=1}^{J} \varepsilon_{ji}(t-1)} EE(t-1) \right) \right\} \cdot \Delta ln\bar{x}_{it}$$

$$+ \frac{1}{2} \sum_{j=1}^{J} \left(\frac{\varepsilon_{ji}(t)}{\sum_{j=1}^{J} \varepsilon_{ji}(t)} EE(t) + \frac{\varepsilon_{ji}(t-1)}{\sum_{j=1}^{J} \varepsilon_{ji}(t-1)} EE(t-1) \right) \cdot \Delta ln\tilde{x}_{jit} \qquad (88)$$

where $ln\bar{x}_{it} = \frac{1}{J} \sum_{j=1}^{J} lnx_{jit}$ and $ln\tilde{x}_{jit} = lnx_{jit} - ln\bar{x}_{it}$. . A similar output mix effect can be obtained if we decompose the output growth in equation (86) taking into account the *average* change in outputs.

5. Concluding remarks

This paper serves as guide to efficiency evaluation from an econometric perspective. The analytical framework relies on the most general parametric models and up to date representations of the production technology through Translog and Quadratic distance functions. We conclude this paper emphasizing the importance of choosing a suitable analytical framework that is in accordance with the industry characteristics and the restrictions faced by the firm, most particularly the relative discretion that managers have over output production and input usage. This sets the stage for the economic objective of the firm that often is assumed to maximize profits (profitability) or minimize cost. Once the theoretical foundation for the measurement of overall economic efficiency is determined, the next question that scholars face is the choice of methods that are available to study variability in firm performance. We discuss the main characteristics, pros and cons, and relevant assumptions that need to be made to successfully undertake a study using SFA techniques.

The extent to which results obtained with the methods surveyed in this paper differ is a general matter of concern that has been addressed by several authors, who employing the same datasets resort to compare the similarity of the distributions of the efficiency scores (see, e.g., Cummins and Zi, 1998). Ultimately, what matters is the ability to provide reliable results on individual performance, not only for the managers of the firms operating within an industry, but also for stakeholders and government agencies involved in regulation, competition and general policy analysis. In this sense, Bauer et al. (1998) propose a set of consistency conditions for the efficiency estimates obtained using alternative methodologies. The consistency of results is related to: 1) the comparability of the estimates obtained across methods, assessed with respect to the efficiency levels (comparable means, standard deviations, and other distributional properties, rankings, and identification of best and worst firms; 2) the degree to which results are consistent with reality, determined in relation to their stability over time, accordance with the competitive conditions in the market, and finally, 3) similarity with standard non-frontier measures of performance. In general, the higher the consistency of efficiency results across all these dimensions, the more confidence regulators and competition authorities will have on the conclusions derived from them, and the intended effect of their policy decisions.

We thus conclude emphasizing the relevance of the methods surveyed in this paper in unveiling the economic performance of firm in terms of technical (and allocative) efficiencies.

Many challenges are still ahead, but cross fertilization of ideas with other research fields will result in a better understanding of the ultimate causes and consequences of inefficient economic performance.

References

- Aigner, D. J., Lovell, C. A. and Schmidt, P. (1977) "Formulation and Estimation of Stochastic Frontier Production Functions," *Journal of Econometrics,* 6:1, 21-37.
- Álvarez, A., Amsler, C., Orea, L. and Schmidt, P. (2006) "Interpreting and Testing the Scaling Property in Models where Inefficiency Depends on Firm Characteristics," *Journal of Productivity Analysis,* 25, 201-212.
- Amsler, C., Prokhorov, A. and Schmidt, P. (2016) "Endogeneity in Stochastic Frontier Models," *Journal of Econometrics,* 190:2, 280-288.
- Amsler, C., Prokhorov, A. and Schmidt, P. (2017) "Endogenous Environmental Variables in Stochastic Frontier Models," *Journal of Econometrics,* 199, 131-140.
- Areal, F. J., K. Balcombe and R. Tiffin (2012), "Integrating spatial dependence into stochastic frontier analysis," *Australian Journal of Agricultural and Resource Economics* 56(4), 521-541.
- Battese, G. E. and Coelli, T. J. (1995) "A Model for Technical Inefficiency Effects in a Stochastic Frontier Production Function for Panel Data," *Empirical Economics* 20:1, 325-332.
- Bauer, P.W., Berger, A.N., Ferrier, G.D. and Humphrey, D.B. (1998) "Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods" *Journal of Economics and Business,* 50:2, 85-114.
- Caudill, S.B., Ford, J.M. and Gropper, D.M. (1995) "Frontier Estimation and Firm-Specific Inefficiency Measures in the Presence of Heteroscedasticity," *Journal of Business,* 13, 105-111.
- Chambers, R. and Quiggin, J. (2000) *Uncertainty, Production, Choice and Agency: The State-Contingent Approach,* New York: Cambridge University Press.
- Chambers, R. G., Chung, Y. and Färe, R. (1996) "Benefit and Distance Functions," *Journal of Economic Theory,* 70, 407-419.
- Chambers, R. G., Chung, Y. and Färe, R. (1998) "Profit, Directional Distance Functions and Nerlovian Efficiency," *Journal of Optimization Theory and Applications,* 95:2, 351-364.
- Chavas, J.P. (2008) "A Cost Approach to Economic Analysis under State-Contingent Production Uncertainty," American *Journal of Agricultural Economics,* 90:2, 435-446.
- Coelli, T., and Perelman, S. (1996) "Efficiency Measurement, Multiple-output Technologies and Distance Functions: with Application to European Railways," No. DP 1996/05. CREPP.
- Colombi, R., Kumbhakar, S., Martini, G., and Vittadini, G. (2014) "Closed-skew normality in stochastic frontiers with individual effects and long/short-run efficiency," *Journal of Productivity Analysis*, 42(2), 123-136.
- Cornwell, C., Schmidt, P. and Sickles, R.C. (1990) "Production Frontiers with Cross-Sectional and Time-Series Variation in Efficiency Levels," *Journal of Econometrics,* 46:1-2, 185-200.
- Cummins, J.D. and Zi, H. (1998) "Comparison of Frontier Efficiency Methods: an Application to the U.S. Life Insurance Industry," *Journal of Productivity Analysis,* 10, 131-152.
- Diewert, W.E. (1971) "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," *Journal of Political Economy,* 79, 461-507.
- Emvalomatis, G. (2009) "Parametric Models for Dynamic Efficiency Measurement," unpublished thesis.

- Emvalomatis, G., Stefanou, S.E. and Lansink, A.O. (2011) "A Reduced-Form Model for Dynamic Efficiency Measurement: Application to Dairy Farms in Germany and the Netherlands," *American Journal of Agricultural Economics,* 93:1, 161-174.
- Färe, R., Grosskopf, S. and Margaritis, D. (2008) "Efficiency and Productivity: Malmquist and more" in Fried, H., Lovell, C.A. and Schmidt, S.S. (eds.), *The Measurement of Productive Efficiency and Productivity Growth,* New York: Oxford University Press.
- Färe, R., Grosskopf, S., Noh, D.W., and Weber, W. (2005) "Characteristics of a Polluting Technology: Theory and Practice*," Journal of Econometrics*, 126, 469-492.
- Filippini, M., and Greene, W. (2016) "Persistent and Transient Productive Inefficiency: A Maximum Simulated Likelihood Approach," *Journal of Productivity Analysis*, 45(2), 187-196.
- Filippini, M., and Greene, W. (2016) "Persistent and Transient Productive Inefficiency: A Maximum Simulated Likelihood Approach", *Journal of Productivity Analysis*, 45:2, 187-196.
- Glass, A. J., Kenjegalieva, K., and Sickles, R. C. (2016) "A spatial autoregressive stochastic frontier model for panel data with asymmetric efficiency spillovers," *Journal of Econometrics*, 190(2), 289-300.
- Greene, W. (1990) "A Gamma-distributed Stochastic Frontier Model," *Journal of Econometrics*, 46:1-2, 141-164.
- Greene, W. (2005) "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model," *Journal of Econometrics*, 126, 269-303.
- Griffiths, W.E., and Hajargasht, G. (2016) "Some Models for Stochastic Frontiers with Endogeneity", *Journal of Econometrics,* 190:2, 341-348.
- Griffiths, W.E., O'Donnell, C.J. and Tan- Cruz, A. (2000) "Imposing Regularity Conditions on a System of Cost and Factor Share Equations," *Australian Journal of Agricultural and Resource Economics,* 44:1, 107-127.
- Guan, Z., Kumbhakar, S.C., Myers, R.J. and Lansink, A.O. (2009) "Measuring Excess Capital Capacity in Agricultural Production," *American Journal of Agricultural Economics,* 91, 765-776.
- Jondrow, J., Lovell, C.A., Materov, S. and Schmidt, P. (1982) "On the Estimation of Technical Efficiency in the Stochastic Frontier Production Function Model," *Journal of Econometrics,* 19:2-3, 233-238.
- Just, R. E. and Pope, R. D. (1978) "Stochastic specification of production functions and economic implications," *Journal of Econometrics*, 7, 67-86.
- Karakaplan, M. U., & Kutlu, L. (2013). Handling endogeneity in stochastic frontier analysis: a solution to endogenous education cost frontier models. Working paper, Department of Economics.
- Kumbhakar, S. C. (2002) "Specification and Estimation of Production Risk, Risk Preferences and Technical Efficiency," *American Journal of Agricultural Economics,* 84, 8-22.
- Kumbhakar, S.C. and Lovell, C.A. (2000) *Stochastic Frontier Analysis*, Cambridge: Cambridge University Press.
- Kumbhakar, S.C. and Tsionas, E.G. (2006) "Estimation of Stochastic Frontier Production Functions with Input-oriented Technical Efficiency," *Journal of Econometrics,* 133:1, 71-96.
- Kumbhakar, S.C., Asche, F. and Tveteras, R. (2013) "Estimation and Decomposition of Inefficiency when Producers Maximize Return to the Outlay: an Application to Norwegian Fishing Trawlers," *Journal of Productivity Analysis*, 40, 307-321.

- Kumbhakar, S.C., Hung-Jen, W. and Horncastle, A.P. (2015) *A Practitioner´s Guide to Stochastic Frontier Analysis Using Stata*, Cambridge: Cambridge University Press.
- Kumbhakar, S.C., L. Orea, L., A. Rodríguez-Álvarez, A. and M. Tsionas, E.G. (2007) "Do We Have to Estimate an Input or an Output Distance Function? An Application of the Mixture Approach to European Railways," *Journal of Productivity Analysis* 27:2, 87-100.
- Kutlu, L. (2010) "Battese-Coelli Estimator with Endogenous Regressors," *Economic Letters*, 109, 79-81.
- Lai, H.P. and Huang, C.J. (2010) "Likelihood ratio tests for model selection of stochastic frontier models," *Journal of Productivity Analysis*, 34: 3-13.
- Malikov, E., Kumbhakar, S.C. and Tsionas, M.G. (2015) "A Cost System Approach to the Stochastic Directional Technology Distance Function with Undesirable Outputs: the Case of US Banks in 2001-2010," *Journal of Applied Econometrics*, 31:7, 1407-1429.
- Meeusen, W. and Van Den Broeck, J. (1977) "Efficiency Estimation from Cobb-Douglas Production Functions with Composed Error," *International Economic Review*, 18:2, 435-444.
- O'Donnell, C. J. and Coelli, T.J. (2005) "A Bayesian Approach to Imposing Curvature on Distance Functions," *Journal of Econometrics, 126*:2, 493-523.
- O'Donnell, C. J. and Griffiths, W.E. (2006) "Estimating State-Contingent Production Frontiers," *American Journal of Agricultural Economics*, 88:1, 249-266.
- O'Donnell, C. J., Chambers, R. G. and Quiggin, J. (2010) "Efficiency Analysis in the Presence of Uncertainty," *Journal of Productivity Analysis*, 33:1, 1-17.
- Olson, J. A., Schmidt, P., and Waldman, D. M. (1980) "A Monte Carlo study of estimators of stochastic frontier production functions," *Journal of Econometrics*, 13(1), 67-82.
- Orea, L. and Kumbhakar, S. (2004) " Efficiency Measurement Using Stochastic Frontier Latent Class Model," *Empirical Economics*, 29:1, 169-183.
- Orea, L. and Wall, A. (2012) "Productivity and Producer Welfare in the Presence of Production Risk," *Journal of Agricultural Economics*, 63:1, 102-118.
- Orea, L., (2002) "A Parametric Decomposition of a Generalized Malmquist Productivity Index," *Journal of Productivity Analysis*, 18, 5-22.
- **Orea, L.**, Álvarez, I. and Jamasb, T. (2018), "A Spatial Stochastic Frontier Model with Omitted Variables: Electricity Distribution in Norway", *Energy Journal*, 39(3), 93-116.
- Orea, L., and Jamasb, T. (2017) "Regulating heterogeneous utilities: A new latent class approach with application to the Norwegian electricity distribution networks," *Energy Journal*, 38(4), 101-128.
- Orea, L., Llorca, M. and Filippini, M. (2015) "A New Approach to Measuring the Rebound Effect Associated to Energy Efficiency Improvements: An Application to the US Residential Energy Demand," *Energy Economics*, 49, 599-609.
- Parmeter, C.F. and Kumbhakar, S.C. (2014) "Efficiency Analysis: A Primer on Recent Advances," *Foundations and Trends in Econometrics*, 7:3-4, 191-385.
- Rodríguez-Álvarez, A. and Lovell, C. A. (2004) " Excess Capacity and Expense Preference Behavior in National Health Systems: an Application to the Spanish Public Hospitals," *Health Economics*, 13:2, 157-169.
- Schmidt A.M., A.R.B. Moreira, S.M. Helfand and T.C.O. Fonseca (2009) "Spatial stochastic frontier models: accounting for unobserved local determinants of inefficiency," *Journal of Productivity Analysis* 31, 101–112.
- Schmidt, P., and Sickles, R. C. (1984) "Production frontiers and panel data," *Journal of Business and Economic Statistics*, 2(4), 367-374.

- Serra, T., Oude Lansink, A. and Stefanou, S. E. (2011)"Measurement of Dynamic Efficiency: A Directional Distance Function Parametric Approach," *American Journal of Agricultural Economics*, 93:3, 756-767.
- Shephard, R.W. (1970) *Theory of Cost and Production Functions*, Princeton, New Jersey: Princeton University Press.
- Silva, E. and Oude Lansink, A. (2013) "Dynamic Efficiency Measurement: a Directional distance function approach. Centro de Economia e Finanças da UPorto, cef.upworking paper 2013-07.
- Simar L., Lovell, C.A. and van den Eeckaut, P. (1994) "Stochastic Frontiers Incorporating Exogenous Influences on Efficiency," Discussion paper no. 9403, Institut de Statistique, Université Catholique de Louvain.
- Stevenson, R. E. (1980) "Likelihood Functions for Generalized Stochastic Frontier Estimation," *Journal of Econometrics*, 13:1, 57-66.
- Thompson, G. D. (1988) "Choice of Flexible Functional Forms: Review and Appraisal," *Western Journal of Agricultural Economics*, 13:2, 169-183.
- Tovar, B. and Wall, A. (2014) "The Impact of Demand Uncertainty on Port Infrastructure Costs: Useful Information for Regulators?," *Transport Policy*, 33, 176-183.
- Tran, K.C. and Tsionas, E.G. (2015) "Endogeneity in Stochastic Frontier Models: Copula Approach without External Instruments," *Economics Letters*, 133, 85-88.
- Tsionas, E. G. (2002) "Stochastic frontier models with random coefficients," *Journal of Applied Econometrics*, 17(2), 127-147.
- Tsionas, E.G. (2006) "Inference in Dynamic Stochastic Frontier Models," *Journal of Applied Econometrics*, 21:5, 669-676.
- Tsionas, E.G. and P.G. Michaelides (2016) "A spatial stochastic frontier model with spillovers: evidence for Italian regions," *Scottish Journal of Political Economy* 63(3), 243-257.
- Wang, H.J. (2002) "Heteroscedasticity and Non-Monotonic Efficiency Effects of a Stochastic Frontier Model," *Journal of Productivity Analysis*, 18:3, 241-253.
- Wang, H.J. (2003) "A stochastic frontier analysis of financing constraints on investment: The case of financial liberalization in Taiwan," *Journal of Business and Economic Statistics* 21, 406-419.
- Wang, H.J. and Ho, C.W. (2010) "Estimating Fixed-Effect Panel Stochastic Frontier Models by Model Transformation," *Journal of Econometrics*, 157:2, 286-296.
- Wang, H.J. and Schmidt, P. (2002) "One-Step and Two-Step Estimation of the Effects of Exogenous Variables on Technical Efficiency Levels," *Journal of Productivity Analysis*, 18, 129-144.
- Wang, W.S. and P. Schmidt, P. (2009) "On the Distribution of Estimated Technical Efficiency in Stochastic Frontier Models," *Journal of Econometrics*, 148, 36-45.
- Winsten, C. B. (1957) "Discussion on Mr. Farrell's paper," *Journal of the Royal Statistical Society*, 120(3), 282-284.
- Zellner, A. and Revankar, N.S. (1969) "Generalized Production Functions," *Review of Economics and Statistics*, 36:2, 241-250.

Appendix

Assume that we have estimated the following multi-input multi-output distance function:

$$\ln D = \ln D\left(x, y, \hat{\beta}\right) \tag{A.1}$$

In order to examine relevant features of firms' technology we should first notice that they must be computed once we assume that the observation belongs to the frontier, i.e. that D=1. Next, we must differentiate the distance function taking into account that $dD = 0$ as we are moving over the frontier. After some simple manipulations we get:

$$0 = \frac{dD}{D} = \sum_{j=1}^{J} \frac{\partial D}{\partial x_j} \cdot \frac{x_j}{D} \cdot \frac{dx_j}{x_j} + \sum_{m=1}^{M} \frac{\partial D}{\partial y_m} \cdot \frac{y_m}{D} \cdot \frac{dy_m}{y_m} \tag{A.2}$$

or

$$0 = \sum_{j=1}^{J} \varepsilon_{Dj} \cdot dlnx_j + \sum_{m=1}^{M} \varepsilon_{Dm} \cdot dlny_m \tag{A.3}$$

where $\varepsilon_{Dj} = \partial lnD / \partial lnx_j$ and $\varepsilon_{Dm} = \partial lnD / \partial lny_m$. The elasticity of output *m* with respect to input *j* can be computed once we assume above that $dlnx_k = 0 \; \forall k \neq j$ and $dlny_n = 0 \;\; \forall n \neq m$, that is:

$$0 = \varepsilon_{Dj} dlnx_j + \varepsilon_{Dm} dlny_m \tag{A.4}$$

This yields the following expression for this specific elasticity:

$$\varepsilon_{mj} = \frac{dlny_m}{dlnx_j} = -\frac{\varepsilon_{Dj}}{\varepsilon_{Dm}} \tag{A.5}$$

Notice that the above elasticity can be computed from both input, output and directional distance function. A return to scale measure (RTS) can be obtained using a similar fashion. In this case, we are interested in the proportional change in outputs caused by a proportional change in all inputs. This implies that $dlnx_j = dlnx \; \forall j = 1, \ldots, J$, and $dlny_m = dlny \; \forall m = 1, \ldots, M$.

$$0 = \sum_{j=1}^{J} \varepsilon_{Dj} dlnx + \sum_{m=1}^{M} \varepsilon_{Dm} dlny \tag{A.6}$$

This yields the following expression for the RTS measure:

$$RTS = \frac{dlny}{dlnx} = -\frac{\sum_{j=1}^{J} \varepsilon_{Dj}}{\sum_{m=1}^{M} \varepsilon_{Dm}} \tag{A.7}$$

Again, the above scale elasticity can be computed from both input, output and directional distance function. However, it can be simplified if we take into account that they satisfy the corresponding homogeneity or transition properties. For instance, if an input distance function has been estimated, $\sum_{j=1}^{J} \varepsilon_{Dj} = 1$, and hence the RTS in (A.7) is equal to:

$$RTS = -\left(\sum_{m=1}^{M} \varepsilon_{Dm}\right)^{-1} \tag{A.8}$$

If instead an output distance function has been estimated, $\sum_{m=1}^{M} \varepsilon_{Dm} = 1$, and hence the RTS in (A.7) collapses to:

$$RTS = -\sum_{j=1}^{J} \varepsilon_{Dj} \tag{A.9}$$