ECONOMIC DISCUSSION PAPERS

Efficiency Series Paper 5/2017

A primer on the theory and practice of efficiency and productivity analysis

Luis Orea, José L. Zofío

Departamento de Economía

Universidad de Oviedo

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

A PRIMER ON THE THEORY AND PRACTICE OF EFFICIENCY AND PRODUCTIVITY ANALYSIS[*]

Luis Orea [a,b]
Oviedo Efficiency Group
Departamento de Economía
Universidad de Oviedo

José Luis Zofío [a]
Oviedo Efficiency Group
Departamento de Análisis Económico: Teoría Económica e Historia Económica
Universidad Autónoma de Madrid

ABSTRACT

This paper introduces the theory and practice of benchmarking the efficiency and productivity of firms, and examines common methodological and empirical choices that researchers face regardless of whether they are performing non-parametric or parametric frontier analyses. We identify different decision forks that researchers encounter, and provide guidance on the options and sequence of steps that should be adopted in order to successfully undertake research in the field. We first summarize the main results of duality theory underlying economic benchmarking, and outline the most popular empirical methods available to undertake efficiency and productivity analyses: DEA and SFA. Afterwards, we discuss several strategies aimed at reducing the dimensionality of the analysis, present a series of models aiming to control for environmental (contextual) variables and endogenous regressors, and discuss the choice of orientation when assessing firms' efficiency using economic criteria. Subsequently we deal with the issue of enhancing the analysis to account for undesirable attributes, such as risk, or proper detrimental outputs like pollutants, waste, contaminants, etc. We next move on to present alternative definitions of temporal productivity change and their decomposition into several terms, such as efficiency change, technical change, scale effects, etc. Finally, dynamic efficiency measurement is discussed.

Keywords: Efficiency and productivity, DEA and SFA, methodological choices, duality, environmental variables, endogeneity, dynamic behaviour.

JEL codes: C4, C5, C6, D24.

---

[*] Parts of the manuscript will be published as a chapter in the forthcoming Palgrave *Handbook of Economic Performance Analysis*.

[a] Both authors wish to acknowledge helpful comments to an earlier version from Juan Aparicio, Bert Balk, Jorge Galan, Magdalena Kapelko, Alfons Oude Lansik, and Alan Wall. Emma Zapico and Joanna Bashford provided efficient research assistance at different stages of this work.

[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).

Contents

A PRIMER ON THE THEORY AND PRACTICE OF EFFICIENCY AND PRODUCTIVITY ANALYSIS

Luis Orea and José Luis Zofío

"Once you choose, it is path-dependent"

1. Introduction

When undertaking a study, scientists must state clearly the methods they follow to answer the postulated research questions. Besides motivating them, they discuss the underlying reasoning as to why particular methods were used and justify their appropriateness. The choice of methods lies at the core of academic work, and reviewing its relevance is natural to any study. This paper is concerned with the choice of methods related to the theory and practice of economic efficiency and productivity analysis. It can be regarded as an opening contribution that sets the stage for the remainder of the volume. Arguably, we contend that a methodological text such as this which serves as a practitioner's guide to undertake research, is mandatory in volumes covering the state of art of a discipline. Here, we provide a general overview of the theory and methods in the field, constituting a qualified index of subsequent contributions which later focus on specific topics.

The concept of scientific research cannot be dissociated from the need to choose, and scholars must weigh the alternatives they have at their disposal to successfully reach their goals. Successful research requires thought processes by which one selects a logical choice from the available options. But to the extent that choosing always entails an opportunity cost in terms of forgone options, it pays to devote enough time to make the right choices. This is because once one settles for an option, investing time in learning the tools of the trade not only constitutes a sunk cost, but creates path dependency, and realizing that the research design was faulty in the final stages of a study, or that the means to a particular end were not adequate, must be avoided. However, it is not always the case that a best option exists among the available alternatives. As in the case of multi-criteria decision making, no preferred option may exist if none dominates over the whole array of attributes defining them. Moreover, there may be alternative ways to approach a research question, with outcomes changing between both, and requiring checks of the robustness and sensitivity of results.

The main theme in this paper is to examine common theoretical and empirical choices that researchers need to select regardless of whether they are carrying out non-parametric or parametric frontier analyses. To address the above-mentioned issues, in this paper we focus on the different forks that researchers encounter, and provide guidance on the options and sequence of steps that should be adopted in order to successfully undertake research in the field. This ranges from the selection of the appropriate economic model to the use of the empirical techniques best suited to achieving reliable results.

Nevertheless, 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. Although we only deal with a subset of available material, the topics considered for discussion are the most recent in the literature, and the references given should be regarded as indicative rather

than exhaustive, granted that we do not aim at an all-encompassing coverage, but to draw attention to the most relevant issues related to the economic side of organizations' performance.

One particular feature of economic efficiency and productivity analyses is that managerial issues are indistinctly addressed by academics affiliated to both business and engineering schools, with the former normally relying on economic science to guide their research questions and methods, and the latter focusing mainly on benchmark measurement and process improvement strategies, pertaining to management science. The distinction is less marked when referring to research methods, as both mathematical programming in operations research and regression analysis in econometrics are empirical approaches that can be used indistinctly to address the same research questions. However, the dividing line between both approaches can be traced by considering whether observations are market oriented or not. Roughly speaking, when market decisions are involved, prices are key in the analysis, and we are in the realm of business and economics. Conversely, if the optimization of production processes and physical operations is the main driver (e.g., production, logistics, supply chain, etc.), then engineering methods and practices prevail. Clearly, this does not rule out studying problems where decision makers do not face market prices explicitly, as would be the case of the provision of non-market oriented services by government agencies and non-profit organizations, given that one could rely on social welfare functions with implicit shadow prices.

But mainly, as declared in the title, this handbook is concerned with the economic side of management practice, exceeding the engineering issues related to production processes, and making it natural to consider the firm, as the relevant decision unit, operating within the market. Therefore, assessing the performance of firms must be undertaken not only from a technical perspective, by checking whether they are capable of producing without incurring input excesses or output shortfalls, but also from an allocative viewpoint, summarized by their capacity to demand and supply optimal amounts of inputs and outputs. Ultimately, it is the ability to consistently abide by the rationality underlying optimal economic behaviour, which ultimately determines the likelihood of long-term survival in the market. This justifies the theoretical focus of the paper on the concept of overall economic efficiency, which starts in the following Section 2 by summarizing the main results of duality theory; particularly the possibility of characterizing the behaviour of the firm from the primal−technological−or dual−economic−perspectives. 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. The existence of dual relationships between particular distance functions (input, output, directional or generalized), and their supporting economic functions (cost, revenue, profit and profitability), offers the researcher the possibility of choosing the perspective of the firms that is best suited for the analysis.

Subsequently, once the theoretical framework to study economic efficiency has been decided, the next question relates to the choice of the most suitable method to characterize the production technology economic behaviour, and, finally, measure firms' performance. Section 3 outlines the most popular empirical methods available to undertake efficiency and productivity analyses; namely non-parametric data envelopment analysis, DEA, and parametric stochastic frontier analysis, SFA. Given space restrictions in terms of text, we limit our discussion to the choice of methods for calculating and estimating primal representations of firms' technology, consistent with the preceding choice of theoretical model. We only identify those key issues related to imposing alternative technological assumptions and properties that, depending on the economic

objective of the firm, require alternative specifications; e.g., quantity and price conditions of homogeneity. When relevant, from this section on we continue a discussion of the methodological and empirical issues which are a matter of concern for practitioners using both approaches.

Section 4 deals with the management of dimensionality difficulties in empirical research; an issue that extends to the choice of flexible functional forms when representing technology or economic behaviour, which demand a large number of observations and information on quantities and prices for consistent estimation. The increased availability of large micro datasets including many variables, and the dimensionality problems associated with low degrees of freedom when the number of observations is limited, compromise the reliability of results and reduce the discriminatory power of the efficiency analysis—particularly in DEA. Consequently, we discuss several strategies aimed at reducing the dimensionality of the analysis, either by relying on dimension reduction techniques that aggregate the original data into a smaller set of composites, or by selecting variables that better characterize production and economic processes. The choice between supervised or unsupervised methods, and the increased information demanded by the former is also considered, along with new proposals. As already mentioned, this section also addresses the choice of functional form when representing the technology, either by using radial, directional or generalized distance functions, each one related to their economic function counterpart.

Another critical issue in the literature, discussed in Section 5, is the need to control for environmental or contextual *z*-variables that do not fall within managerial discretion. The methods used need to control for these characteristics given that they influence individual firm's performance in technical and allocative terms, and in doing so thus ensure a level playing field for all observations (e.g., in regulated industries, local demand characteristics and individually regulated prices). In DEA this has been accomplished with one-step formulations or two-step procedures based on regressions methods for truncated data (e.g., Tobit regressions). Both approaches lead to different results when explaining the effect of contextual variables on efficiency, and therefore it is mandatory to know their specificities. The inclusion of environmental variables in SFA also followed a two-step method initially. However, the pitfalls associated to model misspecification due to the exclusion of relevant variables in the first stage were immediately clear. To deal with this issue, non-discretionary variables have both been included as frontier regressors or as determinants of firms' inefficiency and we discuss the implications that choosing each option has for researchers, managers and policy makers.

Also, the fact that some variables may be endogenous or exhibit a large correlation with firms' inefficiency is gaining increasing attention in the literature. Although the consequences in terms of the reliability of results might prove severe for specific applications, few researchers have tested the independence (or lack-of-correlation) hypothesis between the regressors and inefficiency. Section 6 presents a series of recent models addressing this issue in the DEA and SFA approaches. In DEA recent progress has been made, and although the methods are involved, Monte Carlo experimentation shows promising results. Nevertheless, unless the necessary algorithms are implemented in the existing software, the complexity of the implementation will certainly discourage practitioners. In the parametric approach, recent contributions propose alternative empirical strategies adapting current regression methods to the SFA framework. 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. Therefore, researchers can choose between several options to deal

with endogeneity. We summarize the main features of these methods, and identify their relative advantages and disadvantages. In this section we also focus on the endogenous nature of the directional distance function when assessing firms' efficiency. In the DEA approach we compare endogenous choices based on economic criteria with alternative exogenous orientations normally used in the literature. Within the SFA approach, we show that direction can be imposed on a particular Translog or quadratic specifications through their corresponding almost-homogeneity and translation invariance conditions, respectively. For both approaches we also present data-driven models that allow identifying individual directions based on local proximity (comparability) criteria.

Section 7 deals with the issue of enhancing the analysis to account for attributes that have been modelled in the literature as undesirable, such as risk, or proper detrimental outputs like pollutants, waste, contaminants, etc. Much progress has been made in the SFA approach to incorporate risk and uncertainty, by modelling stochastic technology and economic behaviour−which is in turn not to be confused with the stochastic term related to the joint error including inefficiency. 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. In this section we summarize several approaches proposed in the applied literature to take these factors into account, these extending well beyond the initial and simple production functions with heteroskedastic error terms representing risk, and hence offering a fuller picture of firms' performance in situations of production and demand uncertainty. Subsequently, we shift our attention to the current debate on how to model undesirable outputs, where recent progress has been made. Scholars are still deliberating on the axiomatic characterization of technology, and the need to jointly model desirable and undesirable outputs and their physical relationship. Notably in this context trade-offs exist in the form of engineering—marginal rates of transformation, represented by shadow prices−which sometimes use the idea of consistency with the materials balance principle as a benchmark. DEA and SFA eco-efficiency models defining ratios of economic value added to an index of environmental indicators are also surveyed.

While overall economic efficiency constitutes the main analytical framework of the paper, a straightforward extension of the distance function is productivity (-change) analysis. Section 8 relies on this representation of the technology to present alternative definitions of temporal productivity change based on them. We study the popular concepts of the Malmquist productivity index and Luenberger productivity indicator and their decomposition into several terms explaining productivity change (efficiency change, technical change…), as well as their relationship to traditional−price or value based−definitions such as those of Fisher and Bennet. This link can be built upon by relying on economic theory approach to index numbers and the exactness between the former indices and the specific−flexible−functional aggregators presented in Section 4; i.e., the Translog and Quadratic functions. We stress that among the 'superlative' indices, a preferred or best definition does not emerge, as their appropriateness depends on the criteria chosen. Mainly the latter consists in the number of tests that they satisfy from an axiomatic perspective, or the underlying flexible aggregator to which they are associated. We also discuss the decomposition of profit and profitability change into quantity and price indices. The former related to the Luenberger (profit change) and Malmquist (profitability change) productivity formulations, and the latter associated to allocative efficiency terms. We close this section highlighting that new developments in the measurement of environmental efficiency translate into the so-called Malmquist-Luenberger productivity indices, the numerical interpretation of which remains open.

Before concluding the paper, the last Section 9 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 and benchmarking 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, together with an associated rule for forming expectations with respect to future input prices and technological advances−e.g., time trends. For each model we refer to suitable estimation methods such as GMM or Bayesian techniques with the stochastic frontier approach, and then go on to deal with econometric issues such as unobserved heterogeneity. Also, non-parametric models allowing for dynamic DEA optimization under adjustment costs are presented. The directional input distance function and their dual cost support have been employed in the literature to enable a full decomposition of dynamic efficiency into technical and allocative sources. Recent proposals to estimate a deterministic parametric specification using the quadratic formulation of the directional distance function are also considered.

2. Theoretical background: firms' objective and decision variables

2.1. Theoretical framework: distance functions, economic behaviour, duality and efficiency

The point of departure of any theoretical and empirical study of efficiency and productivity is whether it has or not an economic dimension. That is, if it is merely concerned with technical performance from an engineering perspective (the ability of observations to produce the maximum amount of outputs to inputs), or if input demands and output supplies are optimal given market prices.[1] 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). While technological (shadow) prices may be obtained, quite often no market benchmark exists against which the researcher can contrast the actual economic behaviour of the observations (an exception is the market for CO_2 emissions). In this case we presume that the objective of the firm is technological, based on quantities only.

On the contrary, as would be the case of firms in an industry, if market prices for inputs and outputs are available, these being common to all observations if a price taking market structure prevails, or different across them as a result of idiosyncratic features such as the existence of local markets, then we can extend our engineering analysis to the firm's market environment. This is the realm of economic theory. One can then determine the firm's overall economic efficiency and, subsequently, decompose it in accordance with technical and allocative criteria. In this case we

[1] Nevertheless, the search for optimal *extrema* is at the core of any benchmarking theory based on measurement—including economic theory. This encompasses production theory, as observations aiming to yield the most output with the least feasible amount of inputs.

presume that the objective of the firm is economic and its analysis requires data on quantities and prices.

If an economic efficiency analysis is planned, we must approach the technical (primal) dimension of the problem from a backward induction perspective, so as to determine the optimal representation of the technology in the first place. If the economic (dual) problem of the firm at hand is cost minimization because output levels are given or are exogenous, then an input orientated representation of the technology such as Shephard's input distance function is necessary. Conversely, if what concerns our study is revenue maximization, an output orientation should be chosen. These two possibilities do not exhaust all available choices as firms normally exhibit a maximizing profit or profitability (return-to-dollar) behaviour.

Indeed, what drives the choice of orientation from a technological perspective is duality theory, which enables us to relate a technical (primal) representation of the technology (i.e., the distance function) with a supporting (dual) economic function, and thereby provides a consistent decomposition of economic efficiency into a technical efficiency term and a (residual) counterpart corresponding to allocative efficiency. Next we outline some of the duality theory. More details can be found in Färe and Primont (1995), who summarize the theory of the firm through duality from an input (cost) and output (revenue) perspective, while Chambers et al. (1996, 1998) introduce duality between the profit function and the directional distance function, and Zofío and Prieto (2006) focus on the relationship between the profitability (return-to-dollar) function and the generalized distance function proposed by Chavas and Cox (1999).[2]

2.2. The multi-output, multi-input production technology: Distance functions

The initial step consists in the characterization of the technology set: $T=\left\{(x,y): x\in\mathbb{R}_{+}^{N}, y\in\mathbb{R}_{+}^{M}, x \text{ can produce } y\right\}$ where x is a vector of input quantities, y is a vector of output quantities, and N and M are the number of inputs and outputs. Equivalent representations of the technology are the input requirement set $L(y)=\{x:(x,y)\in T\}$ and the output production possibility set $P(x)=\{y:(x,y)\in T\}$.[3] If the technology exhibits constant returns to scale CRS, then the corresponding set is denoted by $\hat{T}=\{(\psi x,\psi y):(x,y)\in T,\psi>0\}$.[4] For the single output case:

[2] The directional distance function by Chambers et al. (1996) corresponds to the concept of shortage function introduced by Luenberger (1992). Luenberger defined a so-called shortage function (Luenberger, 1992; p. 242, Definition 4.1), which measures the distance of a production plan to the boundary of the production possibility set in the direction of a vector *g*. In other words, the shortage function measures the amount by which a specific plan falls short of reaching the frontier of the technology. Chambers et al. (1996) redefine the shortage function as efficiency measure, introducing the concept of directional distance function.

[3] Based on Debreu's (1951) 'coefficient of resource utilization', Aparicio et al. (2016) introduce the concept of a *loss distance function* generalizing previous representations of production technology, and go on to identify the minimum conditions necessary to derive a dual relationship with a supporting economic function. Duality theory requires that T is a non-empty, closed, convex, and bounded set, with freely disposable inputs and outputs. They obtain the specific normalizing set of the loss function corresponding to the most usual or standard distance functions, thereby nesting previous approaches.

[4] In empirical studies approximating the technology through DEA, the global CRS characterization is assumed for analytical convenience because relevant definitions such as profitability efficiency and the Malmquist productivity

M = 1, the technology can be represented in what is termed as the primal approach by the production function f: $\mathbb{R}_+^N \to \mathbb{R}_+$, defined by: $f(x): \max\{y:(x,y)\in T\}$; i.e., the maximum amount of output that can be obtained from any combination of inputs. The advantage of this interpretation is that it leaves room for technical inefficiency, since under the appropriate assumptions we can define a technology set parting from the production function as $T = \{(x,y): f(x) \geq y, y \in \mathbb{R}_+\}$. Nevertheless, in the general (and real) multiple output-multiple input case, a suitable representation of the technology is given by the distance function introduced by Shephard (1953, 1970).

This representation can be made from alternative orientations including the following distance functions, DF:

- The input DF: $D_I(x,y) = \max\{\lambda:(x/\lambda,y)\in T\} = \max\{\lambda:(x/\lambda)\in L(y)\}$, (1)
- The output DF: $D_O(x,y) = \min\{\theta:(x,y/\theta)\in T\} = \min\{\theta:(y/\theta)\in P(x)\}$, (2)
- The hyperbolic DF: $D_H(x,y) = \min\{\varphi:(x\varphi, y/\varphi)\in T\}$, (3)
- The directional DF: $D_T(x,y;-g_x,g_y) = \max\{\beta:(x-\beta g_x, y+\beta g_y)\in T\}$, (4)
- The generalized DF: $D_G(x,y;\alpha) = \min\{\delta:(\delta^{1-\alpha}x, y/\delta^{\alpha})\in T\}$. (5)

If the technology satisfies the customary axioms, the *output* distance function 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. In contrast, the *input* distance function 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. The *hyperbolic* distance function inherits its name from the hyperbolic path that it follows towards the production frontier. As noted in Section 7, it has the virtue of treating desirable and undesirable outputs asymmetrically. The range of the hyperbolic distance function is $0 \leq D_H(x,y) \leq 1$. It satisfies the following properties: it is almost homogeneous (Aczél, 1966, Chs.5,7; Lau, 1972), non-decreasing in outputs and non-increasing in inputs.

More recent and flexible characterizations are the additive *directional* distance function and the *multiplicative generalized* distance function. The directional distance function, DDF, is a measure of the maximal translation of (x,y) in the direction defined by $g = (g_x, g_y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M \setminus \{0_{N+M}\}$ that keeps the translated input-output combination inside the production possibility set—in (4) and what follows, unless otherwise stated, we denote $g = (-g_x, g_y)$ to emphasize that inputs are reduced. The multiplicative *generalized* distance function, GDF, rescales (x, y) according to the parameter $0 \leq \alpha \leq 1$, also keeping the projected input-output combination inside the production possibility set. These functions nest Shephard's input and output distance functions depending on the specific values of the directional vector $g \neq 0$, or directional parameter α—see Chambers et al. (1996, 1998) and Chavas and Cox (1999). Additionally, the GDF is the only one which nests the hyperbolic distance function for α=0.5. This implies that both approaches can generalize Shephard's input and output distance functions, and therefore their inverse technical efficiency measures correspond to Farrell's (1957) radial efficiency definitions.

index require CRS, and therefore their associated distance functions are defined with respect to that benchmark technology.

However, we note in what follows that such a generalization for the case of the DDF, does not extend to the notion of economic (cost or revenue) efficiency and its decomposition into technical and allocative efficiency, given that the latter does not verify a dual relationship when characterizing the technology through the DDF as shown by Aparicio, Pastor and Zofío (2017).

Therefore, the choice of direction by the researcher, addressed in Section 6.2, represents an initial challenge. Settling for an input or output orientation restricts the production or economic analysis to one dimension ($(-g_x, g_y) = (-x, 0)$, α=0, and $(-g_x, g_y) = (0, y)$, α=1, respectively), while allowing for alternative directions requires justification, including those that assign different directions for each observation.[5] We remark that although the aforementioned distance functions rely on the same set of variables, and sometimes share the same parametric representation, the question that naturally arises is which formulation should be used in DEA and SFA applications.

2.3. Optimizing economic behaviour

We introduce here some alternative economic objectives in order to discuss the duality framework allowing for an overall economic efficiency analysis. Based on the previous primal representations of the technology (1)–(5), and considering the vectors of input and output prices, $w \in \mathbb{R}^N_+$ and $p \in \mathbb{R}^M_+$, the following economic functions can be defined:

- The cost function: $C(y,w) = \min_x \{wx : x \in L(y)\}$, (6)
- The revenue function: $R(x,p) = \max_y \{py : y \in P(x)\}$, (7)
- The profit function: $\pi(w,p) = \max_{x,y} \{py - wx : (x,y) \in T\}$, (8)
- The profitability function (RD): $\rho(p,w) = \max_{x,y} \left\{ py / wx : (x,y) \in \hat{T} \right\}$, (9)
- The shadow cost function: $C(y,w^s) = \min_x \left\{ w^s x : x \in L(y) \right\}$. (10)

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. Correspondingly, the revenue function represents the maximum possible revenue of using a given amount of inputs, yielding the output supply function. The profit function is the maximal feasible profit defined as revenue minus cost−Hotelling's lemma also applies, while the profitability or return-to-dollar (RD) function represents the maximum attainable revenue to cost ratio. It is also possible to define shadow economic functions 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 shadow price vector w^s in (10) rationalizes the observed input quantity vector x as a cost-minimizing choice for the observed output vector y. If the

[5] Daraio and Simar (2016) present the most comprehensive discussion on the different alternatives for the choice of directional vector $(-g_x, g_y)$, ranging from a common or "egalitarian" vector, to one accounting for the heterogeneity of the observations and their diverse contexts that may influence their input and output mixes. A similar approach could be considered for the directional parameter α. A discussion of the different models addressing the choice of orientation is undertaken in section 6.2.

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 behaviour on the part the bureaucrat, restricted by a budget constraint ($wx \leq P$).

The particular properties satisfied by the distance functions with respect to inputs and outputs homogeneity, as well as by the economic functions (related to nonnegativity, monotonicity, homogeneity, convexity and continuity in prices), can be found in standard textbooks, e.g., Mas-Colell et al. (1995), and we shall recall them when needed. Here we highlight that for the optima (min or max) to exist, additional technological conditions for the aforementioned axioms are required. In the case of the profit function, non-increasing returns to scale are required, either equalling 0 or $+\infty$ under constant returns to scale. For the profitability function, Zofío and Prieto (2006) prove that maximum profitability is attained in *loci* where the production technology exhibits local constant returns to scale—i.e., processes exhibiting a technically optimal scale, Balk (1998: p. 19); and constituting a most productive scale size, MPSS, in Banker and Thrall's (1992) terminology. Consequently, a suitable definition of the generalized distance function intended to measure overall economic efficiency must be defined for a production possibility set allowing for constant returns to scale (i.e., using as a benchmark the virtual—cone—technology, $\hat{T}$).[6] For the shadow cost function, see the Orzechowski (1977), Grosskopf and Hayes (1993), and Rodríguez-Álvarez and Lovell (2004).

2.4. Duality and overall economic efficiency: technical and allocative efficiency

Several authors, including Shephard (1970) for (6)-(7), Chambers et al. (1998) for (8) and Zofío and Prieto (2006) for (9) prove the duality between the aforementioned distance functions and their associated economic functions. Based on their interpretation as efficiency measures characterizing production technology (representation property), it is possible to measure overall efficiency and learn about the possible technical and allocative sources of inefficiency. In this respect, we obtain the following dualities: [7]

- Cost: $L(y) = \left\{x : w\left(x / D_I(x,y)\right) \geq C(y,w)\right\}$, (11)
- Revenue: $P(x) = \left\{y : p\left(y / D_O(x,y)\right) \leq R\left(x,p\right)\right\}$, (12)
- Profit: $T = \left\{(x,y) : py - wx + D_T(x,y;-g_x,g_y)(pg_y + wg_x) \leq \pi\left(w,p\right)\right\}$, (13)
- Profitability: $\hat{T} = \left\{(x,y) : \left(p\left(y\,\hat{D}_G(x,y;\alpha)^{-\alpha}\right)\right) / \left(w\left(x\,\hat{D}_G(x,y;\alpha)^{1-\alpha}\right)\right) \leq \rho\left(w,p\right)\right\}$. (14)

---

[6] The technology may be characterized by variable returns to scale as in (3), allowing for scale (in) efficiency $\hat{D}_G(x,y;\alpha) = D_G(x,y;\alpha) \cdot$ SE, with SE = $\hat{D}_G(x,y;\alpha) \,/\, D_G(x,y;\alpha)$, but the final technological benchmark corresponds to constant returns to scale.

[7] Here we are taking into account that $L(y) = \left\{x : D_I(x,y) \geq 1\right\}$, $P(x) = \left\{y : D_O(x,y) \leq 1\right\}$, $T = \left\{(x,y) : D_G(x,y;\alpha) \leq 1\right\}$, and $T = \left\{(x,y) : D_T(x,y,-g_x,g_y) \geq 0\right\}$. For the case of the profit and directional distance functions, the additive overall efficiency measure is normalized by the condition $(pg_y + wg_x)$, ensuring that it is independent of the measurement units as its multiplicative counterparts—see Nerlove (1965).

These dual relations are economic particularizations of Minkowski's (1911) theorem: every closed convex set can be characterized as the intersection of its supporting halfspaces. In fact, the cost, revenue, profitability and profit functions are known as the support functions associated with their corresponding set (see Rockafellar, 1972; p. 112).[8] From an economic point of view, this theorem allows establishing the following decompositions of overall economic efficiency:

- Overall cost efficiency: $\frac{C(y,w)}{wx} = \frac{1}{D_I(x,y)} \cdot AE_I$, (15)
- Overall revenue efficiency: $\frac{py}{R(x,p)} = D_O(x,y) \cdot AE_O$, (16)
- Overall profit (Nerlovian) inefficiency: $\frac{\pi(w,p) - (py - wx)}{pg_y + wg_x} = D_T(x,y;-g_x,g_y) + AI_T$, (17)
- Overall profitability (RD) efficiency: $\frac{py/wx}{\rho(w,p)} = \hat{D}_T(x,y;\alpha) \cdot \widehat{AE}_T$. (18)

The above relationships constitute the core of the analysis of productive and economic efficiency.[9] Several remarks are relevant for applied research. First note that for the overall profitability decomposition, the constant returns to scale benchmark characterizes the generalized distance function. Secondly, a less restrictive property, homotheticity, is also required for a meaningful decomposition of overall economic efficiency, where the distance functions can be rightly interpreted as measures of technical efficiency. Within a non-parametric and parametric setting, Aparicio et al. (2015) and Aparicio and Zofío (2017) respectively show that, for non-homothetic technologies, the radial contractions (expansions) of the input (output) vectors resulting in efficiency gains do not maintain allocative (in)efficiency constant along the firm's projection to the production frontier (isoquants). This implies that they cannot be solely interpreted as technical efficiency reductions. From the perspective of, for example, the cost and revenue efficiency decomposition, this result invalidates the residual nature of allocative efficiency, and requires the use of a distance function with a directional vector capable of keeping allocative efficiency constant along the projections.[10] Thirdly, while the additive DDF nests the input and output radial distance functions for $(-g_x, g_y) = (-x, 0)$ and $(-g_x, g_y) = (0, y)$, respectively, such generalization does not extend to the notion of cost or revenue efficiency and its decomposition

[8] For example, in terms of Aparicio et al.'s (2016) loss distance function $L(x,y;NS)$ measuring the distance from (x,y) to the (weakly) efficient frontier of the technology T—calculated in terms of the normalization set *NS* defined over the price vectors (w,p), the directional distance function (4) defines as: $L(x,y;NS) = \inf_{(w,p)\in R_+^{N+M}} \{\pi(w,p) - (px - wx) : wg_x + pg_y = 1\} = \inf_{(w,p)\in R_+^{N+M}} \{\pi(w,p) - (py - wx)/(wg_x + pg_y)\} = \vec{D}_T(x,y;-g_x,g_y)$, where $wg_x + pg_y = 1$ is the associated normalizing condition.

[9] Two other particular cases of the loss distance function $L(x,y;NS)$ dual to the profit function are the Hölder distance function (Briec and Lesourd, 1999) and the weighed additive distance function (Aparicio et al., 2016). Nevertheless, they are by far less popular than their DDF counterpart and have been implemented only in the non-parametric DEA approach.

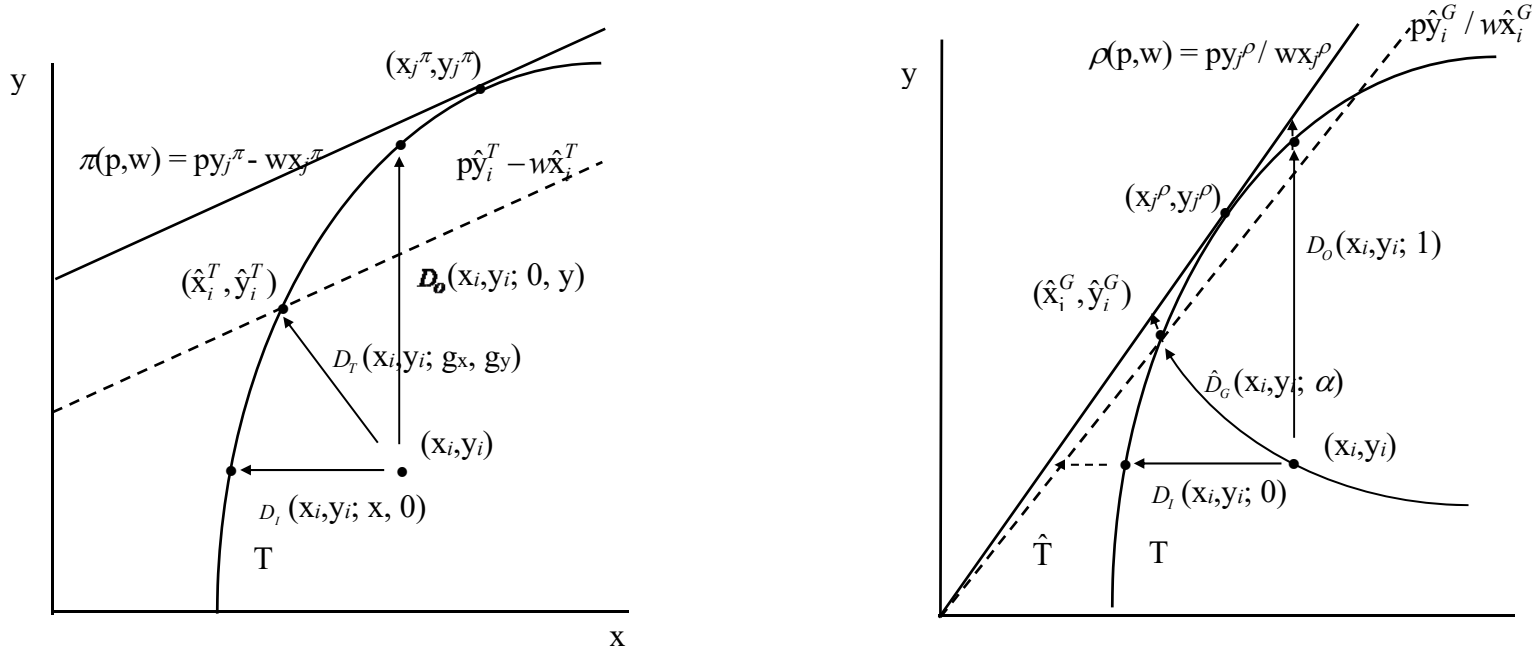
[10] This in turn implies that the radial framework or choosing as a directional vector the observed amounts of inputs and outputs in the case of the directional distance function is no longer valid.

into technical and allocative terms. For these particular directions, allocative efficiency cannot be obtained as an independent residual from the above inequalities as shown by Aparicio, Pastor and Zofío (2017). [11]

The alternative distance function representations of production technology (technical efficiency measures), dual economic functions, and residual nature of allocative efficiency are presented in Figure 1. We illustrate overall economic efficiency for the most general cases corresponding to the directional and generalized distance functions, and their dual profit and profitability functions. In the left panel (a) the directional function (4) measuring the distance from the single input-single output unit (x_i,y_i) to the frontier is represented by $D_T\left(x_i, y_i; g_x, g_y\right)$, measuring technical inefficiency and, equivalently—thanks to the duality relationship (13)—its associated profit and loss in monetary units if the normalizing constraint is set to $(pg_y + wg_x) = 1$. This projects the unit to point $(\hat{x}_i^T, \hat{y}_i^T)$, whose profit is $p\hat{y}_i^T - w\hat{x}_i^T$. Therefore, and thanks to (17), the difference between maximum profit—attained at (x_i^π, y_i^π)—and observed profit, corresponds to allocative inefficiency: $\pi(p,w) - (p\hat{y}_i^T - w\hat{x}_i^T)$. In the same panel (a) the input and output distance functions (1) and (2) are also presented as particular cases of the directional formulation for $(-g_x, g_y) = (-x, 0)$ and $(-g_x, g_y) = (0, y)$, but whose interpretation in terms of overall cost or revenue efficiency is inconsistent. The right panel (b) presents an equivalent analysis in terms of the generalized distance function (5) projecting the evaluated unit to $(\hat{x}_i^G, \hat{y}_i^G)$ through $\hat{D}_G(x, y; \alpha)$. In the single input-single output case its technical inefficiency interpretation is the amount by which observed average productivity y_i / x_i, can be increased to attain, $\hat{y}_i^G / \hat{x}_i^G$ at the reference frontier. Now, thanks to the duality relationship (14), the difference can be interpreted in terms of profitability differentials given the input and output market prices. Finally, following (18), it is possible to determine allocative efficiency as the ratio between projected profitability $p\hat{y}_i^G / w\hat{x}_i^G$ to maximum profitability—attained at (x_i^ρ, y_i^ρ): i.e., $(p\hat{y}_i^G / w\hat{x}_i^G) \,/\, py_i^\rho / wx_i^\rho$. Again, in the same panel (b) the input and output distance functions are presented for α=0 and α=1, respectively.

[11] Finally, we note that a firm is overall profit efficient when its technical and allocative terms are zero rather than one. This implies that the larger the numerical value of the directional distance function the more inefficient is the firm, thus the technical and allocative (*in*)efficiency notation*:* *TI* and *AI*, with $TI = D_T(x, y; -g_x, g_y)$. Balk (1998) favours a consistent characterization of efficiency throughout, so the larger the value the greater the firm's efficiency. This is achieved by multiplying (17) by minus one, resulting in $TE = -D_T(x, y; -g_x, g_y)$.

Figure 1. Distance functions and their economic duals: Profit (a) and profitability (b).



3. Empirical background: Standard approaches to measure firms' economic efficiency

Once the basic theoretical framework regarding the choice of economic model is presented, the next step is the consideration of the empirical methods that allow the measurement of firms' efficiency using both production (primal) and economic (dual) approaches. Particularly, since both the true technology and economic behaviour of the firms are unknown, they are often approximated by using either non-parametric mathematical programming, parametric econometric techniques (regression analysis), or engineering (bottom-up) models. In this section we only describe the main features of the two most popular approaches, DEA and SFA. Throughout the section we limit our discussion to simple specifications of both parametric and non-parametric models.

3.1. Data Envelopment Analysis

Following Koopmans (1957), DEA approximates production technology from observed historical, cross-sectional, or panel data relying on the activity analysis approach and mathematical programming. Based on the principle of minimum extrapolation, DEA yields the smallest subset of the input-output space $\mathbb{R}_+^N \times \mathbb{R}_+^M$ as an inner approximation containing all observations, and satisfying certain technological assumptions. Technology consists of piecewise linear combinations of the observed $i = 1,.., I$ firms constituting a multidimensional production frontier.[12] The DEA piecewise linear approximation of the technology T, is given by

$$T = \left\{ (x,y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M : \sum_{i=1}^{I} \lambda_i \, x_{in} \leq x_n, n = 1,...,N; \sum_{i=1}^{I} \lambda_i \, y_{im} \geq y_m, m = 1,...,M; \; \sum_{i=1}^{I} \lambda_i \; = 1, \lambda \in \mathbb{R}_+^I, i = 1,...,I \right\}, \tag{19}$$

where λ is an intensity vector whose values determine the linear combinations of *facets* which define the production frontier and whose associated restrictions allow considering alternative returns to scale. Among the technology axioms incorporated into the above DEA model we highlight convexity, strong disposability, and variable returns to scale.

Regarding *convexity*, while there are alternative DEA models dropping this assumption like the Free Disposal Hull (FDH) or Free Replicability Hull (FRH), these are inconsistent with duality theory (*i.e.*, Minkowski's theorem), since convexity is key when recovering the supporting economic functions. As for *free (or strong) disposability*, implying that it is feasible to discard unnecessary inputs and unwanted outputs without incurring in technological opportunity costs, is a rather weak assumption that, nevertheless, has its drawbacks. Most importantly, when measuring technical efficiency through radial distance functions, their values reflect whether the firm belongs to the so-called isoquant subsets (e.g., Isoq $L(y)$ or Isoq $P(x)$), a rather weak notion of inefficiency, that leaves room for non-radial efficiency improvements associated to strong disposability; i.e., weak disposability is required for the distance function to characterize the production technology

[12] See Cooper, Seiford and Tone (2007) and Färe, Grosskopf and Lovell (2008) for an introduction to the Activity Analysis DEA within a production theory context.

as in (11)–(14). Indeed, relying on the notion of Pareto efficiency, it is accepted that efficiency is to be measured against the efficient subset of the production possibility set:

$$Eff(T) = \{(x,y) \in T : (u,-v) \leq (x,-y), (u,v) \neq (x,y) \Rightarrow (u,v) \notin T\}. \quad (20)$$

Much research effort has been devoted to propose DEA models that measure efficiency against this subset, or incorporate the slacks associated to a weak (non Pareto) definition of the efficiency set:

$$WEff(T) = \{(x,y) \in T : (u,-v) < (x,-y) \Rightarrow (u,v) \notin T\}. \quad (21)$$

Therefore, there is a trade-off between the axiom of strong disposability of inputs and outputs (and their corresponding DEA implementation), resulting in technologies with weakly efficient subsets, against which radial and other distance functions with preassigned directions cannot account for possible slacks.[13] Regardless of the disposability assumption, it is possible to formulate DEA programs that ensure efficiency measurement against the efficient subset. This is the case of the additive models despite the strongly disposable characterization of the production technology.[14] Finally, alternative *returns to scale* can be postulated in (19) through the intensity variables λ. The variable returns to scale assumption can be dropped in favour of constant returns to scale by removing $\sum_{i=1}^{I} \lambda_i = 1$, and non-increasing and non-decreasing returns to scale correspond to $\sum_{i=1}^{I} \lambda_i \leq 1$ and $\sum_{i=1}^{I} \lambda_i \geq 1$, respectively.

Once the technology is defined, it is possible to calculate the distance functions (1)−(5) by solving their corresponding mathematical programs. As the input, output and hyperbolic distance functions are particular cases of the generalized and directional distance functions here, we present these latter formulations. Taking as a guiding framework the decomposition of economic efficiency, and the associated scale properties of the technology associated to profit and profitability maximization, corresponding to non-increasing and constant returns to scale, we consider the following programs in order to evaluate the efficiency of firm $(x_{i'}, y_{i'})$:

[13] This trade-off has prompted research on the general problem of transforming any weak DEA (in)efficiency measure into a strong DEA (in)efficiency, e.g., Fukuyama and Weber (2009) and Pastor and Aparicio (2010). Pastor et al. (2016) show that *any* DEA model that projects inefficiency observations onto the weakly efficient frontier, rather than onto the strongly efficient frontier, can be related to a reversed directional distance function, RDDF. They propose a two stage process that combines a given efficiency measure, which offers a first stage projection for each observation on the weak efficient frontier, with a second stage additive model that projects each first stage projection onto the strongly efficient frontier, ending up with a strongly efficient projection. Relating each inefficient observation with that final second stage projection through the corresponding RDDF, results in a comprehensive DDF (in)efficiency measure that combines radial and non-radial inefficiencies into a single scalar.

[14] There is a specific axiom for *joint production* across outputs, which is a realistic assumption in environmental efficiency studies abiding by the second law of thermodynamics stating that production without waste is impossible, and therefore strong disposability must be dropped. In this case the technology satisfies: (i) weak disposability, by which a reduction in desirable, good or market outputs can only be achieved with simultaneous and proportionate reductions of the unintended, undesirable or bad outputs; and (ii) null-jointness (no emission generation implies no production of market outputs). We discuss the specific axioms of environmental models in section seven below.

Directional Distance Function, *DDF*

$$
\begin{aligned}
& D_T^s(x_{i'}^t, y_{i'}^t; -g_x, g_y) = \\
& = \max_{\beta,\lambda_i} \left\{ \beta : (x_{i'}^t - \beta g_x, y_{i'}^t + \beta g_y) \in \mathrm{T}^s \right\} \\
& \text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i \, x_{in}^s \le x_{i'n}^t - \beta g_{x_n}, n = 1,...,N, \\
& \qquad \sum_{i=1}^{\mathrm{I}} \lambda_i \, y_{im}^s \ge y_{i'm}^t + \beta g_{y_m}, m = 1,...,M, \\
& \qquad \sum_{i=1}^{\mathrm{I}} \lambda_i = 1, \quad \lambda \in \mathbb{R}_+^I.
\end{aligned}
\tag{22}
$$

Generalized Distance Function, *GDF*

$$
\begin{aligned}
& \hat{D}_G^s\left(x_{i'}^t, y_{i'}^t; \alpha\right) = \\
& = \min_{\delta,\lambda_i} \left\{ \delta : (x_{i'}^t \delta^{1-\alpha}, y_{i'}^t / \delta^{\alpha}) \in \hat{\mathrm{T}}^s \right\} \\
& \text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i \, x_{in}^s \le \delta^{1-\alpha} \, x_{i'n}^t, \quad n = 1,...,N, \\
& \qquad \sum_{i=1}^{\mathrm{I}} \lambda_i \, y_{im}^s \ge y_{i'm}^t / \delta^{\alpha}, \quad m = 1,...,M, \\
& \qquad \lambda \in \mathbb{R}_+^I,
\end{aligned}
\tag{23}
$$

which incorporate the DEA production possibility set presented in (19). Note that the notation has been enhanced to include time superscripts identifying time periods: *s*, *t*, to which the technology and the firms are referred, with *s*, *t*= 0, 1 representing the base and comparison period, respectively. This notation allows for both contemporary and "mix-period" distance functions that are necessary to compute the Malmquist and Luenberger productivity definitions presented in Section 8.4.[15]

Besides the values of the distance functions representing the technical efficiency scores, relevant information can be obtained both from the above "envelopment" formulations of the technology and their "multiplier" duals. As identified in (22) and (23), an advantage of DEA is that it yields explicit, real-life benchmarks. A firm can indeed study the technical and economic behaviour of its peers, so as to improve its own efficiency and productivity−those firms with optimal $\lambda_i^* > 0$ conform the reference frontier, and the value corresponds to the relevance of the benchmark firm in the linear combination. Regarding technical efficiency, the number of possible peer firms is equal to the number of inputs plus the number of outputs except in the CRS case, where there can generally be one less reference peer. This follows from linear programming theory in that there exists an optimal solution for which the number of positive variables is at most equal to the number of linear restrictions.

Second, from the dual of the above programs, technological relationships between inputs and outputs can be discerned, in the form of shadow prices involving their multipliers (ν, μ) and defining the supporting (reference) hyperplanes against which technical efficiency is measured. In this case the firm is efficient if it belongs to one of the supporting hyperplanes (forming the facets of the envelopment surface) for which all firms (x_i,y_i) lie on or beneath it. The duals corresponding to the directional distance function and generalized distance functions are the following:[16]

[15] These contemporary or mix-period programs, allowing for a flexible choice of the directional vector and ($-g_x$, g_y) and parameter α can be solved using the DEA Toolbox developed by Álvarez et al. (2016).

[16] The dual for the GDF envelopment formulation (23) can be determined because it corresponds to a CRS characterization of the production technology, rendering it equivalent, for instance to the radially oriented output distance function (2) for $\alpha = 1$−since the value of $\hat{\mathrm{D}}_G(x,y;\alpha)$ is independent of α. As for the VRS counterpart, the non-linear nature of (23) hampers the formulation of the dual "multiplier" program.

Dual Directional Distance Function

$$
\begin{aligned}
&\mathrm{D}_T^s(x_{i'}^h, y_{i'}^h, -g_x, g_y) = \\
&= \min_{\mu,\nu,\omega} \nu x_{i'}^h - \mu y_{i'}^h + \omega \\
&\text{s.t.} \quad -\nu\ x_i^s + \mu^s \mathrm{y}_i^s - \omega \le 0, \quad i = 1,...,I \\
&\qquad \nu\ g_x + \mu\ g_y = 1, \\
&\qquad \nu\ \ge 0,\ \mu\ \ge 0.
\end{aligned} \tag{24}
$$

Dual Generalized Distance Function

$$
\begin{aligned}
&\hat{\mathrm{D}}_T^s\left(x_{i'}^h, y_{i'}^h; \alpha\right) = \max_{\mu,\nu} \mu \quad y_{i'}^h \\
&\text{s.t.} \quad -\nu\ x_i^s + \mu\ \mathrm{y}_i^s \le 0, \quad i = 1,...,I \\
&\qquad \mu\ x_{i'}^h = 1, \\
&\qquad \nu\ \ge 0,\ \mu\ \ge 0.
\end{aligned} \tag{25}
$$

The choice of the primal "envelopment" formulations (22)–(23) or their "multiplier" duals (24)–(25) depends on the analytical objective of researchers and the specific characteristics of the study. Nevertheless, we note that the simplex method for solving the envelopment form also produces the optimal values of the dual variables, and all existing optimization software provides both sets of results readily, so there is not any computational burden on a particular choice of model.[17] For peer evaluation and determination of the nature of returns to scale the envelopment formulations are adequate, while the duals are required if one wants to set weight restrictions rather than to adhere to the "most favourable weights" that DEA yields by default (Thompson et al., 1986; Podinovsky, 2015). Also, as optimal weights are not unique, one can define secondary goals in comparative analyses that, using cross-efficiency methods, also help to rank observations that are efficient in the standard (first stage) DEA (Sexton et al., 1986; Cook and Zhu, 2015).

Once the distance functions measuring technical efficiency have been calculated, it is possible to determine the efficiency of overall cost, revenue, profit and profitability, subject to the same technology. Such programs incorporate the restrictions characterizing the production possibility sets, and jointly determine minimum cost or maximum revenue, profit or profitability, depending on the choice of firm's economic behaviour. The following programs correspond to the profit and profitability cases in a given period $t = 0, 1$, while the overall cost and revenue can be calculated similarly:[18]

[17] Nevertheless, the computational effort of solving the envelopment problems grows in proportion to powers of the number DMUs, *I*. As the number of DMUs is considerably larger than the number of inputs and outputs (*N*+*M*), it takes longer and requires more memory to solve the envelopment problems. We contend that except for simulation analyses and the use of recursive statistical methods such as bootstrapping, nowadays processing power allows calculation of either method without computational burdens.

[18] There are recent contributions decomposing profit and profitability change into quantity and price terms with a technical and allocative efficiency change interpretation−i.e., the former corresponding to the Malmquist index and Luenberger indicator, that require calculation of mix-period programs of profit and profitability efficiency, see Zofío and Prieto (2006) and Juo et al. (2015).

Profit efficiency

$$\frac{\pi^t(p^t,w^t)-(p^t y^t_{i'}-w^t x^t_{i'})}{p^t g^t_y + w^t g^t_x} = \max_{\beta,\lambda_i,x^t,y^t} \quad \varphi$$

$$\text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i \, x^t_{in} = x^t_n, \, n = 1,...,N,$$

$$\sum_{i=1}^{\mathrm{I}} \lambda_i \, y^t_{im} = y^t_m, \, m = 1,...,M, \qquad (26)$$

$$\sum_{i=i}^{\mathrm{I}} \lambda_i \, \frac{py^t_i - w^t x^t_i}{pg^t_y + wg^t_x} = \frac{py^t_{i'} - wx^t_{i'}}{pg^t_y + wg^t_x} + \varphi,$$

$$\sum_{i=1}^{\mathrm{I}} \lambda_i = 1, \quad \lambda \in \mathbb{R}^I_+.$$

Profitability efficiency

$$\frac{p^t y^t_{i'} / w^t x^t_{i'}}{\rho^t(p^t,w^t)} = \min_{\phi,\lambda_i,x^t,y^t} \quad \phi$$

$$\text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i \, x^t_{in} = x^t_n, \, n = 1,...,\mathrm{N},$$

$$\sum_{i=1}^{\mathrm{I}} \lambda_i \, y^t_{im} = y^t_m, \, m = 1,...,\mathrm{M}, \qquad (27)$$

$$\sum_{i=i}^{\mathrm{I}} \lambda_i \, \frac{w^t x^t_i}{p^t y^t_i} = \phi \frac{w^t x^t_{i'}}{p^t y^t_{i'}},$$

$$\lambda \in \mathbb{R}^I_+.$$

The decomposition of overall economic efficiency can then be completed by calculating allocative efficiency as the residual closing equations (15)–(18). The importance of the flexibility of the directional model (26) is stressed by Aparicio et al. (2015) when decomposing overall economic efficiency. These authors show that only in the case of homothetic technologies, can the standard radial measures *a la Farrell* be considered as correct measures of technical efficiency. Otherwise, under non-homotheticity, the standard estimations would measure an undetermined mix of technical and allocative efficiency. To restore a consistent measure of technical efficiency in the non-homothetic case they introduce a method that takes as reference for the economic efficiency decomposition the preservation of the allocative efficiency of firms producing in the interior of the technology. This builds upon the so-called reversed approach recently introduced by Bogetoft et al. (2006) that allows calculating allocative efficiency without presuming that technical efficiency has already been accomplished. They illustrate their method in the non-parametric approach, adopting the simplest non-homothetic variable returns to scale model and illustrate how to implement them with a numerical example using KLEM data. They show that there are significant differences in the allocative and technical efficiency scores between the standard and consistent approaches.

3.2. Stochastic Frontier Approach

In this section we outline the main features of the standard econometric approach to measuring firms' inefficiency. For a comprehensive survey of this literature, see Kumbhakar and Lovell (2000), Fried et al. (2008) and Parmeter and Kumbhakar (2014). For notational ease, we have developed this section for cross-sectional data, except when it is compulsory to use a panel data framework. In analogy to the DEA analysis, we confine our discussion to the estimation of technical efficiency using distance functions. Thus, firm performance is evaluated by means of the following (general) distance function:

$$\ln D_i = \ln D(x_i, y_i, \beta) + v_i, \qquad (28)$$

where the scalar y_{it} is the output of firm i =1,…I, x_i is a vector of inputs, lnD_i measures firm' technical efficiency, and ln$D(x_i,y_i,\beta)$ is a deterministic distance function,[19] β is now a vector of technology parameters, and v_i is a two-sided noise term with zero mean. In equation (28) 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.

A relevant issue that should be addressed here is that while the dual representations of the technology in (6)−(9) are clearly identified in a parametric exercise by the different sets of both dependent and explanatory variables, this is not the case for the primal representations based on the distance functions in (1)−(5). At first sight, all of them are functions of 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? Note also that, regardless the orientation of the distance function, the term measuring firms' inefficiency (i.e. lnD_i) is not observed by the researcher and thus it cannot be used as a proper dependent variable to estimate (28).

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, output and generalized (hyperbolic) 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.[20] Identification works because each homogeneity condition involves different sets of variables. Although the underlying technology is the same, the coefficients of each distance function differ.[21]

In the case of Shephard's output distance function, we have that it is linearly homogenous in outputs. The implication of this property is that, normalizing by one of the outputs, say y_{1i}, the deterministic distance function ln$D(x_i,y_i,\beta)$ can alternatively be rewritten as:

$$\ln D\left(x_i, y_i, \beta\right) = \ln D\left(x_i, y_i / y_{1i}, \beta\right) + \ln y_{1i}. \tag{29}$$

Note that this specification immediately "produces" an observed dependent variable for the above model once (29) is inserted into (28). Indeed, rearranging terms, the model in (28)can be expressed as follows:

$$-\ln y_{1i} = \ln D\left(x_i, y_i / y_{1i}, \beta\right) + v_i - \ln D_i, \tag{30}$$

[19] It is implicitly assumed here that ln$D(x_i,y_i,\beta)$=0 as deviations from this (frontier) value are captured by lnD_{it}, which however is unobserved by the researcher.

[20] In passing we note that the DEA mathematical programs are also defined so as to satisfy the homogeneity conditions of each distance function.

[21] It is worth mentioning that $D(x,y)$ can be viewed as a general specification of firms' technology that nests all distance functions in (1)−(5). This general representation of the technology is equivalent to the general transformation function used in Kumbakahar (2012) and Parmeter and Kumbhakar (2014). These authors show that this function cannot be estimated without imposing specific normalizations on the set of parameters to be estimated. Moreover, Parmeter and Kumbhakar (2014) point out that the econometric estimation will yield different results because of the fact that different exogenous assumptions are (sometimes implicitly) made.

or equivalently as[22]

$$\ln y_{1i} = -\ln D\left(x_i, y_i / y_{1i}, \beta\right) + v_i - u_i, \tag{31}$$

where $u_i = -\ln D_i \geq 0$ is a non-negative random term measuring firms' inefficiency that can vary across firms.

Note that this model can be immediately estimated econometrically once a particular functional form is used for $\ln D(x_i, y_i, \beta)$, and u_i is properly modelled. The input and hyperbolic distance functions, and the directional distance function deserve similar comments.[23] While the Shephard's distance functions are mainly estimated in the literature using the Translog specification, the Quadratic function is often used as a parametric specification for the directional distance function since Chambers (1998) and Färe et al. (2005) showed that this specification offers the advantage that it can be easily restricted to satisfy the translation property. Both Translog and Quadratic functions are not only differentiable allowing for the estimation of shadow prices and output/input substitutability, but also provide a second-order approximation to a true, but unknown distance function (see Section 4.3 below on the choice of flexible functional forms).

Note that the error term $\varepsilon_i = v_i - u_i$ in (31) 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, we assume u_i to be non-negative (and asymmetrically) distributed. This results in a composed error term ε_i that is asymmetrically distributed. Finally, as customary in the literature, we assume in this paper that both random terms are distributed independently of each other and of the input variable.

3.2.1. Estimation methods

We now turn to explaining how to estimate the above frontier model. 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 (30) can first be estimated via *maximum likelihood* (ML) once particular distributional assumptions on both random terms are made. This is the most popular empirical strategy in the literature, but it relies on (perhaps strong) assumptions regarding the distribution of both random terms, and the exogeneity nature of the regressors. Conditional on the ML estimated parameters, efficiency scores can then be estimated for each firm by decomposing the estimated residual into a noise component and an inefficiency component.

---

[22] Note that here we have not changed the sign of the noise term because the distribution of v_i and $-v_i$ is the same as they follow normal distributions.

[23] As these functions are respectively linearly homogenous in inputs and almost homogenous, the corresponding distance functions to be estimated are $\ln D(x,y) = \ln D\left(x/x_1, y\right) + \ln x_1$, and $\ln D(x,y) = \ln D\left(xy_1, y/y_1\right) + \ln y_1$. Regarding the directional output distance functions, the translation property says that if output is expanded by ζg_y and input is contracted by ζg_x, then the resulting value of the distance function is reduced by ζ. If one chooses the neutral orientation $(-g_x, g_y) = (-1, 1)$, the directional distance function is $D_T(x, y, \beta; -1, 1) = D_T(x - \zeta, y + \zeta, \beta; -1, 1) + \zeta$. By choosing a specific ζ coefficient for each firm, a variation on the left hand side is obtained.

A second method that we can choose is the *method-of-moments* (MM) approach. The MM approach involves three stages. In the *first* stage, all technological parameters of the production function are estimated using appropriate econometric techniques (e.g., OLS or GMM if the input variables are exogenous or endogenous, respectively). This stage is independent of distributional assumptions in respect of either error component. Thus the first-stage MM estimates are robust to non-normality and heteroscedasticity of the unknown error term (Verbeek, 2008, p.143). The fact that the MM approach allows for endogenous regressors explains why it is becoming more popular among researchers. 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 (i.e., the variance of v_i and u_i), conditional on the first-stage estimated parameters.[24] In the *third* stage, efficiency scores are estimated for each firm by decomposing the estimated residuals.

The second-stage of the MM approach can also be estimated using the second and third moments of the error term ε_{it} in equation (31) if we follow the so-called *Modified Ordinary Least Squares* (MOLS) method first proposed by Afriat (1972) and Richmond (1974). 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. Most researchers using the MM approach prefer using a ML estimator in the *second* stage of the MM estimation procedure instead of the set of moment conditions because MOLS has some practical problems even in homoscedastic specifications of the model. For instance, the implied σ_u might become sufficiently large to cause $\sigma_v<0$, which violates the assumptions of the econometric theory.

Whatever the approach we favour, we are forced to choose a distribution for v_i and u_i in order to estimate (partially) the parameters in equation (31) by ML. While the noise term is often assumed to be normally distributed with zero mean and constant standard deviation, several distributions have been proposed in the literature for the inefficiency 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 mean zero and constant standard deviation, that is $u_i \sim N^+(0,\sigma_u)$.[25] The most important characteristic of this distribution is that the modal value of the inefficiency term (i.e., the most frequent value) is close to zero, and higher values of u_i are increasingly less likely (frequent). 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.[26] It should be pointed out that the identification of both random terms in these models relies on the one-sided nature of the distribution of u_i and not necessarily on the asymmetry of the

[24] Thus, the above ML approach merely combines the two first stages of the MM approach into one.

[25] 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 u_i. On the other hand, we are modelling the variance (not the mean) of the pre-truncated normal distribution in this specification. It should be taken into account that at the end of the day we are modelling the mean (and the variance) of u_i as it is a function of the variance of the original pre-truncated normal distribution.

[26] Other distributions proposed in the literature are the Pearson distribution of Lee (1983), the uniform distribution of Li (1996), the binomial distribution of Carree (2002), the Beta distribution of Gagnepain and Ivaldi (2002), the Laplace-truncated Laplace distribution of Horrace and Parmeter (2014), or the Cauchy-Half-Cauchy distribution introduced by Nguyen (2010).

inefficiency term (see Li, 1996). In other words, if the inefficiency term could take both positive and negative values, it would not be distinguishable from the noise term.

Let us briefly assume that the standard model (28) has a temporal dimension, with the I firms observed in periods $t = 0,..,T$. Under the normal-half-normal distributional assumptions, the log likelihood function for a sample of I firms can then be written as (see Kumbhakar and Lovell, 2000; p.77):

$$\ln L(\beta,\gamma,\delta) = -\frac{IT}{2}\ln\left(\sigma_v^2+\sigma_u^2\right) + \sum_{i=1}^{I}\sum_{t=1}^{T}\ln\Phi\left(\frac{-\varepsilon_i^t\,\sigma_u/\sigma_v}{\left(\sigma_v^2+\sigma_u^2\right)^{1/2}}\right) - \sum_{i=1}^{I}\sum_{t=1}^{T}\frac{\varepsilon_i^{t2}}{2\left(\sigma_v^2+\sigma_u^2\right)}, \tag{32}$$

where $\varepsilon_i^t = \ln y_{1i}^t + \ln D\left(x_i^t, y_i^t / y_{1i}^t, t, \beta\right)$. The likelihood function (32) can be maximized with respect to (β,γ,δ) to obtain consistent estimates of all parameters of the model.

Several comments are in order regarding the above log likelihood function. First, note that in (32) we will obtain the same likelihood function, and thus the same parameter estimates, if we reverse the summation or we recode our observations *as if* we had a cross-sectional dataset with IT observations in a unique period. That is, the double summation in (32) reveals that our model is a pooled model because it does not distinguish between cross-sectional and temporal observations. In other words, although the above likelihood function uses both types of observations, it is not a panel data model! [27] A related issue that is rarely recognized in stochastic frontier literature is that, although ML estimates based on an incorrect assumption of independence over time still lead to consistent estimates, the standard errors of the estimates, calculated under the assumption of independence, will be wrong if independence does not hold. Fortunately, it is possible to calculate "corrected" estimated variances by using the "sandwich form" expression suggested by Alvarez et al. (2006).[28]

Second, the above distributional assumptions provide closed form solutions for the distribution of the composed error term, making the direct application of ML straightforward. Newer models (see Parmeter and Kumbhakar; 2014, and the references in their Section 7) are appearing in literature that do not yield tractable likelihood functions and must be estimated by simulated maximum likelihood.

Finally, so far we have assumed that the inefficiency and noise terms are independently distributed. This could be a strong assumption for instance in agriculture where noisy and seasonal fluctuations often affect productive decisions. The error components independence assumption has been recently relaxed by Bandyopadhyay and Das (2006) and Smith (2008). While the first paper assumes that v_i^t and u_i^t are jointly distributed as normal truncated bivariate so that u_i^t is truncated

[27] Many young researches confuse having a panel data set and using a panel data estimator. In this sense, it must be noted that the (heteroscedastic) "model for technical inefficiency effects in a stochastic frontier production function for panel data" introduced by Battese and Coelli (1995) is just a pooled model that, despite the title of the paper, does not take advantage of the panel structure of the data because it assumes independence over time. Indeed, our model also assumes that the u_{it} are independent over time. This is widely recognized as an unrealistic assumption, but it is not clear how to relax it. See Parmeter and Kumbhakar (2014, section 7) for an excellent review of the literature dealing with this issue and how to distinguish between persistent and time-varying inefficiency.

[28] Another issue, often overlooked, is that the conditional expectation $E(u_{it}|\varepsilon_{it})$ in a pooled model does not provide consistent estimates of the efficiency scores.

below zero, the second uses the copula approach. The copula allows parameterizing the joint behaviour of v_i^t and u_i^t and tests the adequacy of the independence assumption.[29] The latter author also shows that the distribution of the composed error term can yield wrong skewness problems, making it difficult to estimate the above model by ML. Also, the ML estimator is subject to significant biases when error component dependence is incorrectly ignored.[30]

3.2.2. Efficiency scores

Once the model has been estimated using ML or MM, the next step is to obtain the efficiency estimates for each firm. In both cases, the composed error term is simply $\varepsilon_i=v_i-u_i$. Hence, we can follow Jondrow *et al.* (1982) and use the conditional distribution of u_i given the composed error term ε_{it} to estimate the asymmetric random term u_i. Both the mean and the mode of the conditional distribution can be used as a point estimate of u_i. However, the conditional expectation $E(u_i|\varepsilon_i)$ is by far the most commonly employed in the stochastic frontier analysis literature (see Kumbhakar and Lovell, 2000).

Given the above distributional assumptions, the analytical form for $E(u_i|\varepsilon_i)$ can be written as follows:

$$\hat{u}_i = E(u_i \mid \varepsilon_i) = \tilde{\mu}_i + \tilde{\sigma}\frac{\phi\left(\tilde{\mu}_i/\tilde{\sigma}\right)}{\Phi\left(\tilde{\mu}_i/\tilde{\sigma}\right)}, \tag{33}$$

where $\tilde{\mu}_i = \frac{-\varepsilon_i\sigma_u^2}{\sigma_u^2+\sigma_v^2}$, and $\tilde{\sigma} = \frac{\sigma_v\sigma_u}{\left(\sigma_u^2+\sigma_v^2\right)^{1/2}}$.

Two comments are in order regarding the above point estimate of u_i. First, Wang and Schmidt (2009) show that inference on the validity of the chosen specification of the inefficiency term should not be carried out by simply comparing the observed distribution of $\hat{u}_i$ to the assumed distribution for u_i. 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 tests, however, are sensitive to other misspecifications, e.g., on the normality distribution of the noise term, and most likely the homoscedastic assumption of both the noise and inefficiency terms.

Finally, the choice of a particular distribution for the inefficiency term should not only rely on statistical criteria, but also on the competitive conditions of the markets where the firms are operating. For instance, the above-mentioned distributions allow for the existence of very inefficient firms in the sample, which is an unappealing feature if they are operating in very competitive markets. In these markets, it might be more appropriate to use the double-bounded distribution introduced by Almanidis et al. (2010) that imposes both lower and upper theoretical

[29] Other papers have used the copula method in other types of SFA applications. For instance, Amsler et al. (2014) use copulas to model time dependence in stochastic frontier models; or Carta and Steel (2012) suggest using copulas in modeling multi-output stochastic frontiers.

[30] Using a set of simulation exercises, Simar and Wilson (2010) show that the wrong skewness issue might also appear even when the underlying skewness is the correct one.

bounds on the values of the inefficiency term. Moreover, both the "quiet life" hypothesis[31] and the results of some recent papers[32] providing evidence on the correlation between market power and operational inefficiency suggest that different market equilibrium outcomes might yield different distributions for the inefficiency term. For instance, partial collusion is a reasonable equilibrium in markets with many firms, as coordination among *all* firms is extremely difficult to maintain in this environment. This equilibrium results in a bimodal distribution for the inefficiency term due to the existence of two sets of collusive (likely inefficient) and competitive firms (likely very efficient).[33] In other markets, *all* firms might be involved in a perfect cartel scheme. In such a cartel-equilibrium, while most firms are setting monopoly prices, a small set of more competitive firms could be setting smaller prices due to cheating behaviour.[34] The quiet life of most firms in this framework suggests that they are likely (very) inefficient and that only a few (cheating) firms are more efficient. This implies that firms' inefficiency is likely to be negatively skewed. In this case, the composed error term might have skewness bearing the wrong sign.

3.2.3. Parametric decomposition of economic efficiency

The previous discussion is concerned with the technical side of the firm, and needs to be extended to the overall economic efficiency concept that constitutes our reference analytical framework. Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014) present alternative models that allow the decomposition of cost, revenue and profit efficiency into technical and allocative terms, based on the estimation of either single equations, or a system of equations including optimal input and output quantities, both applicable to cross-section and panel data cases. In this section we discuss relevant contributions that consider one output for simplicity, and can therefore be presented using the production function definition. The extension to the multiple-output, multiple-input technology through directional distance functions has been addressed by Aparicio and Zofío (2017). This can be related to recent developments in duality theory that facilitate the decomposition of overall economic efficiency in the case of homothetic technologies. When technologies are non-homothetic, new results show that the directional distance function becomes the corner stone allowing for the identification and measurement of

[31] As Das and Kumbhakar (2016) point out this hypothesis postulates that lack of competitive pressure reduces managers' effort to seek operational efficiency. In the absence of sufficient disciplining devices, managers may pursue objectives other than profit maximization and they might spend resources in order to obtain (or maintain) market power. As a result, the extra rents earned, thanks to market power, might simply allow inefficient managers to persist.

[32] Huang et al. (2017) uses a copula-based stochastic frontier method to model the correlation between cost efficiency and market power. Delis and Tsionas (2009), Koetter and Poghosyan (2009), Koetter et al. (2012), and Das and Kumbhakar (2016) also estimate market power within a model that uses stochastic frontier analysis, but in these contributions stochastic frontier analysis is used *only* to recover cost inefficiency, and subsequently determine its impact on market power.

[33] This is exactly the situation that tries to capture the so-called Zero Inefficiency Stochastic Frontier (ZISF) model introduced by Kumbhakar et al. (2013). These authors propose using a latent class structure to distinguish between fully efficient firms, and firms that are inefficient to some extent. While the inefficiency distribution for fully efficient firms is a point mass at 0, the degree of inefficiency for inefficient firms is captured by any of the array of standard one-sided distributions, such as half-normal, exponential, or truncated normal. The ZISF model is also appealing in a benchmarking context, as it helps regulators to identify the utilities that can be used as "reference networks" for other (comparable) utilities.

[34] For instance, Ellison (1994) finds that secret price cuts occurred during 25% of the cartel period with price discounts averaging about 20%.

technical and allocative efficiency in a consistent way. Also, new developments within dynamic efficiency models accounting for economic efficiency are considered in Section 9 below.

One of the first proposals to decompose overall economic efficiency parametrically is presented by Kopp and Diewert (1982). These authors introduce a method based solely on duality theory, without resorting to the primal approach with direct or indirect knowledge of a production function, and its associated minimum cost share equations, as previous proposed by Schmidt and Lovell (1979) or Kopp (1981). Instead they simply employ knowledge of the relevant economic function. Considering cost minimization as a reference, a system of equations involving optimal demands for inputs−applying Shephard's lemma−and relative input quantities, allows determination of the unknown reference technical efficient benchmark for any firm as well as its associated input−shadow−price vector. Based on this solution a straightforward decomposition of cost efficiency into technical and allocative terms is possible. Subsequent refinements by Zieschang (1983) and Mensah (1994), who improved the method by simplifying the system of equations to be solved, resulted in less computational requirements and numerical difficulties.

To present this method we follow Aparicio and Zofío (2017) who revisit it for the case of non-homothetic technologies relying on the directional distance function. They show that the process is simplified when the analysis involves a self-dual homogenous technology; i.e., the cost function $C(y,w)$ and production function $f(x)$ can be analytically recovered from each other−as in the Cobb-Douglas or the generalized production function cases. They focus on the input side of the firm and its associated overall cost efficiency(15), but the idea can be extended to revenue and profit efficiencies with the necessary qualifications.

The method can be formally elaborated by recalling the production function representing the technology, which is assumed to be homogenous of degree r: $f(\delta x)=\delta^r f(x)$, $\delta>0$, $r>0$, as well as the cost function (6), $C(y,w)=\min_x\{wx: x\in L(y)\}$, representing the minimum cost of producing y given the vector of inputs prices $w = (w_1, \ldots, w_N)$. In this case the following relationships are verified: $D_I(x,y)=1/D_O(x,y)^{1/r}$, and the cost function is separable, i.e. $C(y,w)=y^{1/r}C(1,w)$.[35]

Given the vector of market prices w one can recover the amount of inputs minimizing the cost of production by way of Shephard's lemma, i.e.,

$$x^*(y,w)=\nabla_w C(y,w), \tag{34}$$

where $\nabla_w C(y,w)\equiv\left[\partial C(y,w)/\partial w_1,...,\partial C(y,w)/\partial w_N\right]$. Again, under degree r homogeneity, the system of demand equations can be expressed as:

$$x^*(y,w)=y^{1/r}\nabla_w C(1,w). \tag{35}$$

[35] Boussemart et al. (2009) introduced a more general definition in the literature for multi-output multi-input contexts: a production technology T is said to be homogeneous of degree α if for all $\lambda>0$ $(x,y)\in T\Rightarrow\left(\lambda x,\lambda^\alpha y\right)\in T$. In particular, if $f(\delta x)=\delta^r f(x)$ for all $\delta>0$ and $T=\left\{(x,y): f(x)\geq y, x\in R_+^N, y\in R_+\right\}$, then T is homogeneous of degree r following Boussemart et al.'s definition.

For any two inputs *n* and *n'* with associated market prices w_n and $w_{n'}$, the first order conditions imply that the marginal rate of technical substitution of factor *n* for factor *n'* must be equal to their price ratio: [36]

$$MRS^{n}_{n'} = -dn'/dn = f_n(x)/f_{n'}(x) = w_n/w_{n'}, \tag{36}$$

where $f_k(x) = \partial f(x)/\partial x_k$, $k = n, n'$, are marginal productivities. With this information we define the allocative efficiency of firm *i*, (x_i, y_i), according to (15), as follows:

$$AE(x_i, y_i, w) = \frac{C(y_i, w)}{w\hat{x}_i} = \frac{C(y_i, w)}{wx_i / D_I(x_i, y_i)} = \frac{w\nabla_w C(y_{it}, w)}{wx_i / D_I(x_i, y_i)} = \frac{wx^*(y_i, w)}{wx_i / D_I(x_i, y_i)}, \tag{37}$$

where hereafter we assume that all input prices are common to all firms as in Kopp and Diewert (1982).

If the firm is allocative efficient, then $AE(x_i, y_i, w)=1$, with $\hat{x}_i = x_i / D_I(x_i, y_i) = x^*(y_i, w)$, and the marginal rates of substitution are equal to the input price ratios at the efficient projection, i.e., (36)is verified. It follows immediately that if $x_i / D_I(x_i, y_i) \neq x^*(y_i, w)$ the firm is allocative inefficient with $AE(x_i, y_i, w) < 1$, with the marginal rates of substitution differing from relative prices.

Decomposing overall cost efficiency (15) is a relatively direct procedure when either the production or cost function is known. In that case the firm's technical efficiency can be directly calculated through $D_I(x_i, y_i)$, or can be recovered from its output counterpart; a simple matter in the single output case since: $D_I(x_i, y_i) = 1/D_O(x_i, y_i)^{1/r} = 1/\left(y_i / f(x_i)\right)^{1/r}$. This allows determining production cost at the efficient projection $w\hat{x}_i = wx_i / D_I(x_i, y_i)$. Finally, from (34) we can recover the optimal input quantities minimizing the cost of producing y_i, $wx^*(y_i, w)$, and following (37) the residual allocative efficiency is the ratio between minimum cost and cost at the efficient projection: $AE(x_i, y_i, w) = wx_i^* / w\hat{x}_i$.

However, when there is only knowledge of the cost function and the primal counterpart cannot be recovered, as it is the case of flexible forms such as the Translog or Quadratic specifications, it is possible to follow the method initiated by Kopp and Diewert (1982), which identifies the technically efficient projections on the isoquant by solving the following system of $2N-1$ equations:

[36] We assume that given our assumptions about production technology, the second order conditions are verified and therefore the sign of the bordered Hessian determinant is negative.

$$\begin{aligned} &\hat{x}_i = \nabla_w C(y_i, w^s), \quad i = 1,...,I, \\ &\hat{x}_{i,n} / \hat{x}_{i,n'} = x_{i,n} / x_{i,n'} \quad , \forall n \neq n', \end{aligned} \tag{38}$$

where w_t^s represents a vector of normalized shadow prices (e.g., using w_1 as numeraire)—see also Balk (1997), and the second set of $N-1$ equations ensures that the input proportions (mix) of the firm under evaluation are kept constant, implying a radial projection that is consistent with the underlying input distance function.

The above procedure assumes that either the production or cost functions are *known*, and it is based on numerical methods that are *deterministic*. However, within the SFA setting, several authors have addressed the decomposition of overall economic efficiency when the estimation of unknown production or cost function is required—for early references see Kumbhakar (1991,1997). Kumbhakar and Lovell (2000) and 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 approach.

As in the deterministic approach above, we present a model that is compatible with the output-oriented technical inefficiency, which under the homogeneity assumption is equivalent to its input-oriented counterpart, in conjunction to the dual cost function.[37] The preferred approach estimates a system consisting of a stochastic production function which allows for technical inefficiency, $f(x_i)e^{v_i-u_i}$ —where, as before, v_i is random noise and $u_i \geq 0$ is the output oriented technical efficiency. By taking logs the first order conditions for cost minimization give the following system of equations:

$$\begin{aligned} &\ln y_i = \ln f(x_i) + v_i - u_i, \\ &\ln s_{ni} - \ln s_{1i} - \ln\left(w_n x_{ni}\right) + \ln\left(w_1 x_{1i}\right) = \xi_{ni}, \quad n = 2,...,N, \end{aligned} \tag{39}$$

where s_{ni} stands for input shares, and the second line is obtained departing from (36) in the following way: $f_n(x_i)/f_1(x_i) = \left(w_n / w_1\right)e^{\xi_{ni}} \Rightarrow \left(\partial \ln f_n(x_i)/\partial \ln x_{ni}\right)/\left(\partial \ln f_1(x_i)/\partial \ln x_{1i}\right) \equiv s_{ni}/s_{1i} = \left(w_{ni}x_{ni}/w_{1i}x_{1i}\right)e^{\xi_{ni}} \Rightarrow \ln s_{ni} - \ln s_{1i} - \ln\left(w_n x_{ni}\right) + \ln\left(w_1 x_{1i}\right) = \xi_{ni}$. The parameter $\xi_{ni} \gtrless 0$ captures allocative efficiency deviations for the input pair (n, 1) —with the sign showing whether input n is over or underused relative to input 1, that serves as numeraire. Departure from the optimality conditions can be measured by the difference in the bilateral ratios corresponding to the marginal productivities and input prices. Considering that firm (x_i, y_i) is technically inefficient: $y_i = f\left(x_i\right)e^{-u_i}$, one can recover the isoquant corresponding to $f(x_i)$ as $y_i e^{u_i}$, which is subsequently used to measure the allocative efficiency of the technically efficient projection of (x_i, y_i). As the production function is homogenous of degree r, the radial output shift of the production isoquants is neutral. In this case output oriented projections leave the marginal rates of substitution (slopes of the isoquants) unchanged, and therefore allocative efficiency is the same regardless of the

[37] Parmeter and Kumbhakar (2014; ch. 4) present the methods for both the input and output oriented approach to technical efficiency, together with the cost and profit functions.

isoquant that is taken as reference for its measurement being either that corresponding to the output amount actually observed $L(y_i)$ or the projected efficient amount $L(y_i e^{u_i})$.

A parametric decomposition of overall cost efficiency solves the system of equations (39) by maximum likelihood for a given functional form (e.g., Translog), and makes the necessary distributional assumptions for the error component as previously discussed: v_i, u_i and ξ_{ni} (i.e., normal, half normal and multivariate normal, respectively). Once again, the error term can be decomposed using the standard procedure proposed by Jondrow et al. (1982), while the allocative inefficiencies ξ_{ni} are obtained from the residuals of the first order conditions in (39). However, the magnitudes of the allocative inefficiencies need to be calculated from the actually observed input quantities and those yielded by the input demands functions (34). Again, the method could be adapted to decompose overall revenue and profit efficiency; for the latter case see Parmeter and Kumbhakar (2014) and Kumbhakar et al. (2015).[38]

As in the case of the non-parametric DEA setting discussed by Aparicio et al. (2015), the previous determinist and stochastic methods are appropriate when the production function is homogenous of degree r, which in the single output case implies homotheticity. Aparicio and Zofío (2017) confirm that, in this case, resorting to radial distance functions to characterize the technology and measure technical efficiency is adequate, since allocative efficiency is independent of the output level and radial input and output shifts leave it unchanged. However, for non-homothetic production functions, they show that the use of radial measures is inadequate because optimal input demands depend on the output targeted by the firm, as does the inequality between marginal rates of substitution and market prices—i.e., allocative inefficiency. They demonstrate that a correct definition of technical efficiency corresponds to the directional distance function, because its flexibility ensures that allocative efficiency is kept constant through movements in the input production possibility set when solving technical inefficiency. Therefore, the associated cost reductions can be solely—and rightly—ascribed to technical improvements—i.e., they endogenize the directional vector (see Section 6.2). Resorting to a deterministic parametric specification they show that it is possible to identify a specific reference vector g_x so that the input directional distance function actually measures technical inefficiency by leaving allocative efficiency unchanged.

Again, the consistency of the new approach is based on the so-called reversed cost efficiency decomposition by Bogetoft et al. (1996), which is equivalent to the Farrell approach in the homothetic case, but yields alternative technical and allocative values when non-homothetic technologies are involved. The latter ensures that once a given output level is considered for the reverse approach, allocative efficiency is first determined, and its technical efficiency counterpart rather than being still associated to the radial input measure, can be correctly determined by calculating the directional vector associated to the directional distance function. The new decomposition for non-homothetic technologies mirrors the properties of the standard approach, including the fact that the technical efficiency for firms situated on a given isoquant is the same when projected to the same reference isoquant. This is thanks to a suitable normalization condition which also ensures that inefficiencies are measured in monetary values. The optimization programs needed to implement the model empirically in a parametric setting are also introduced, these allowing the calculation of the directional distance function and its associated directional vectors

[38] Kumbhakar et al. (2015; chs. 8 and 9) show how to implement these methods using STATA.

as well as illustrating the new methods with homothetic and non-homothetic Cobb-Douglas specifications.

3.3. Evaluating, comparing and reconciling DEA and SFA methods for decision making

Once the basic characteristics of the standard DEA and SFA approaches have been presented, it is clear that the individual results, rankings and distributions obtained from both methods will generally differ. However, the difference between the non-parametric and parametric methods is less pronounced nowadays than they used to be because both approaches now benefit from recent advances that address their shortcomings. In the past, it was their deterministic or stochastic nature, and therefore their relative ability to accommodate noise and error that marked the difference. A second difference was their non-parametric and parametric nature, preventing second order differentiability and proneness to misspecification error, respectively. In passing we note that most DEA results are based on the envelopment programs (22) and (23), which successfully identify reference peers, but do not offer a characterization of the production technology. However, to unveil the characteristics of production technology and economic optima one must resort to the "multiplier forms" (24) and (25) providing linear hyperplanes (facets), from which one gains information regarding shadow prices, marginal productivities, rates of substitution and transformation, etc. Still, the mathematical programming approach does not enjoy the interpretative convenience of the econometric approach; e.g., in terms of average technical and economic elasticities (scale, cost, revenue, etc.). In turn, in this latter approach it is the 'average' firm that characterizes technology, rather than the observations at the frontier, which are essentially those that represent 'best practice' behaviour, i.e., those with the highest efficiency, productivity and economic performance.

From the perspective of DEA, its deterministic nature has been qualified thanks to the extension of statistical methods to mathematical programming. This is the case of chance constrained DEA and bootstrapping techniques based on data resampling, which can be customarily found in several software packages thanks to the increase in processing capacity. As for the need to adopt a specific−even if flexible−functional form in SFA, that may satisfy the desired regularity conditions locally, and be prone to misspecification bias, the availability of semi-parametric and Bayesian techniques is opening new opportunities−e.g., Kumbhakar et al. (2007). Also, new proposals based on convex nonparametric least squares (CNLS) and the so-called Stochastic Nonparametric Envelopment of Data (StoNED), are also trying to bridge the gap between both methods—Johnson and Kuosmanen (2015).

The extent to which results obtained with both approaches differ is a general matter of concern that has been addressed by several authors, who employing the same datasets resort to non-parametric tests to compare the similarity of the distributions of the efficiency scores. One of the first studies comparing the results of alternative methods is Hjalmarsson et al. (1996), who undertake a comparison of DEA scores with the deterministic and stochastic frontier approaches, DFA and SFA, for a panel data of fifteen Colombian cement plants during 1968-1988. Immediately after, subsequent studies by Bauer et al. (1998) on the U.S. banking industry and Cummins and Zi (1998) on US life insurance, not surprisingly, confirm that pairwise comparisons of results within alternative mathematical programming models (e.g., DEA and FDH) or econometric models (e.g, SFA vs. TFA) yield higher correlations than those between methods.

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. Bauer et al. (1998) is a relevant contribution because it proposes a set of consistency conditions for the efficiency estimates obtained using alternative methodologies: The reference DEA and SFA methods, enhanced with Distribution Free Analysis (DFA) and Thick Frontier Approach (TFA). The consistency of results is related to:

1) The comparability of the estimates obtained across methods, assessed with respect to: a) the efficiency levels (comparable means, standard deviations, and other distributional properties), b) rankings, and c) identification of best and worst firms; and

2) The degree to which results are consistent with reality, determined in relation to: a) their stability over time, b) accordance with the competitive conditions in the market, and finally, c) similarity with standard non-frontier measures of performance. In general, the higher the consistency of efficiency results across all these dimension, the more confidence regulators and competition authorities will have on the conclusions derived from them, and the intended effect of their policy decisions.

These authors survey a number of studies on financial institutions and examine all six of these consistency conditions for regulatory usefulness. They also perform an analysis of overall economic efficiency solving all four previous frontier approaches on a single dataset of 683 U.S. banks over a 12-year period. To draw comparable results, they consider the same sets of variables, sample periods (either cross section or panel data), etc. Pairwise comparisons of results showing the consistency within non-parametric and parametric methods has been already mentioned. As for the degree to which they are consistent with reality, yielding credible results, they indicate that all of the methods provide rankings that are stable over time, but SFA and TFA regressions appear to show higher consistency with what are generally believed to be the competitive conditions in banking markets, and also with the most usual non-frontier measures of bank performance, such as return on assets or various cost ratios, which are often used by regulators, managers, and consultants.

When many studies on a particular industry are available, a meta-analysis of the results helps to discern the main stylized facts characterizing the production technology, economic behavior and firms' performance. A meta-analysis yields a weighted average of the different studies performed under alternative assumptions, closer to the true results than those yielded by individual studies. The question is how these weights are allocated across the studies and the way in which the uncertainty is computed around the average estimate that is obtained. Besides providing an estimate of the unknown common truth, this method allows researchers to confront results from different studies and identify patterns among them, the sources of disagreement, and other relevant relationships that may come to light in the context of multiple studies. On these grounds, it represents a suitable method to identify the main factors behind the observed variability in efficiency. Brons et al. (2005) and Bravo-Ureta et al. (2007) study mean technical efficiency values in farming and transport, while Oh and Lee (2010) focus on productivity change using Malmquist indices. Clearly, results from alternative models vary due to several factors including the methods used (e.g., non-parametric vs. parametric), alternative model specifications (e.g., returns to scale), specific observations and variables (e.g., non-discretionary), time periods (e.g., cross-section or panel data), etc. The objective is to determine how efficiency values vary under the different choices, and how robust results are for a particular industry. Also, Odeck and Bråthen

(2012) perform a comprehensive meta-analysis using the Tobit regression—as efficiencies can be interpreted as corner or censored solutions—on forty seaport published studies—twenty-nine using DEA and eleven SFA, where the differences across them are the explanatory variables. They compare a fixed-effects specification implying that a true effect is shared by all the studies, to a random-effects model that allows the true effect to vary among the different studies included in the data set. From the perspective of the methods they find that (i) the random-effects model outperforms the fixed effects model in explaining the variations mean technical efficiencies; (ii) studies relying on non-parametric DEA models yield higher values than SFA models (as expected given their deterministic nature), and (iii) that panel data studies have lower scores as compared with those using cross-sectional data.

4. Some (critical) issues when modelling firms' technology

This section is devoted to the discussion of several issues related with firms' technology or, in other words, with the shape of the deterministic part of the frontier *conditional* on the number of variables, and the number of frontier determinants.

New technologies allow researchers to collect larger amounts of data. With respect to the number of observations and variables, a relative trade-off exists between these which serves to determine the confidence level and usefulness of results. It is summarized within the concept of degrees of freedom which is relevant for both mathematical programming (DEA) and stochastic frontier regression (SFA) perspectives. Degrees of freedom is a function relating the sample size (I) with the number of independent variables (e.g., N inputs and M outputs), and, the larger the number of independent observations with respect to the number of variables, the greater the confidence offered to researchers when making inferences about a statistical population (i.e., hypotheses testing). This serves for both approaches when performing parametric tests relying on asymptotic theory, meaning that theoretical properties can be established for large sample (e.g., regarding parameter estimates, the nature of returns to scale, input and output substitutability, etc.).[39]

Moreover, besides the use of parametric tests, in DEA, the ability of these methods to discriminate observations by their efficiency is compromised when the numbers of observations is limited. As DEA methods search for the most favourable weights for the firm under evaluation, it is more likely to assign weights that render the firm efficient when there are less degrees of freedom. Any firm for which the ratio of outputs to inputs can be maximized by varying weights (including zero values), will be deemed efficient−i.e., when an extreme firm employs the lowest amount of any of the N inputs, or produces the largest amount of any of the M outputs it is going to be categorized as efficient. In a DEA context this situation has resulted in a "rule of thumb"

[39] Alternative hypotheses testing methods corresponding to non-parametric and bootstrap-based inference have been proposed in the literature. In this context, we depart from Cook et al. (2014), who consider that since DEA is not a form of regression model, but rather a frontier-based linear programming optimization technique, it is meaningless to apply sample size requirements, and it should be viewed as a benchmarking tool focusing on individual performance. They suggest that in order to increase the discriminatory power of DEA when the number of firms is relatively low with respect to that of inputs and outputs, one could use weight restrictions or other DEA approaches to reduce the number of efficient units; e.g., see Allen et al. (1997) for a discussion on weight restrictions and value judgement, and Chen (2012) on the TOPSIS approach using ideal and anti-ideal observations.

proposal by which the number of observations should be at least twice the number of inputs and outputs: $I \geq 2(NxM)$−Golany and Roll (1989), while Banker et al. (1989) raise this threshold to three: $I \geq 3(NxM)$. However, if the population is small with industries composed of just a few firms in a particular market, the DEA benchmarking results can still be helpful, while a regression based analysis may yield inconclusive regression results−regardless of the lack of statistical validity.

While the availability of massive datasets has reshaped statistical thinking, and computational power allows carrying out statistical analyses on large size databases, the trade-off between observations and variables persists as a relevant issue in many applications. Reducing the dimensions of data is a natural and sometimes necessary way of proceeding in an empirical analysis using either DEA or SFA. Indeed, dimension reduction and variable selection are the main approaches to avoid the curse of dimensionality.

4.1. Dimension reduction

This empirical strategy can be viewed as a two-stage procedure. In the first stage, a set of variables are aggregated into a small number of composites or aggregated variables. In a second stage, the composites are plugged into a production or cost frontier that is estimated using either non-parametric or parametric techniques. Therefore, this approach reduces the input (output) dimensionality of the data set by replacing a set of decision variables and regressors with a lower-dimensional function.

The most common methods used to achieve this objective are principal component analysis (PCA) and explanatory factor analysis (EFA).[40] The dimensionality of the data set is reduced using these statistical methods by expressing the variance structure of the data through a weighted linear combination of the original variables. Each composite (ordered in decreasing order of percentage variance) accounts for maximal variance while remaining uncorrelated with the preceding composite. These methods have been used to carry non-parametric efficiency analyses. For instance, Ueda and Hoshiai (1997) and Adler and Golany (2001, 2002) develop PCA-DEA models to obtain the efficiency estimates where a set of principal components replace the original variables. Other remarkable applications of this approach are Adler and Yazhemsky (2010) and Zhu (1998). As only a percentage of the information is retained from each of the original variables, the discriminatory power of the DEA model is improved. Yu et al. (2009), Nieswand et al. (2009) and Growitsch et al. (2012) use these PCA and EFA in a SFA framework to control for the effect of many environmental conditions on cost efficiency of electricity distribution networks.

From an analytical perspective, this two-stage procedure implicitly assumes that the technology is separable. Separability hinges on how the marginal rate of transformation between two individual variables only depends on the variables within the composite. Therefore, a necessary condition for the existence of a theoretically consistent composite is the separability of

[40] A detailed discussion of the statistical pros and cons of each method is beyond the scope of this paper. It suffices to note that there is no consensus among statistical theorists as to what conditions should determine the use of PCA or EFA. PCA is often preferred as a method for data reduction, while EFA is often preferred when the goal of the analysis is to detect structure. Moreover, some authors argue that the difference between the two techniques is negligible (e.g., Velicer et al., 1990).

the elements within the aggregate from those outside the aggregate.[41] Otherwise, the use of composites in estimation may well be subject to specification errors. A tentative action is to test the existence of separability using cost and production functions as in Kim (1986). However, when the number of inputs (outputs) is large, the precision of these tests is probably too low to be used with confidence. Moreover, carrying out such tests can be an impossible task when the dimensionality problem becomes truly severe.

Another analytical issue is whether the procedure to aggregate individual inputs (outputs) allows us to get consistent parameter estimates. For example, the theory of index numbers has shown that the appropriate functional form of the aggregate depends on the characteristics of the technology. According to this theory, there is only one appropriate (i.e., "superlative") index or composite for each technological representation (see Diewert, 1976)−see Section 8.3 below. There is no guarantee that both the statistical-based and exact composites coincide with each other. Moreover, Jamasb et al. (2010) show using an error-in-variable model that we should expect the existence of endogeneity problems if the statistical-based composite does not coincide with the exact composite.

4.1.1. Unsupervised dimension reduction techniques

From a statistical point of view, the main drawback of both PCA and EFA is that they might misspecify the fundamental relationship between the dependent variable and the variables to be aggregated because both methods ignore information on the dependent variable when reducing the dimension of the data. In this sense, PCA and EFA are *unsupervised* dimension reduction techniques in the terminology coined by Fisher (1922). A simple theoretical model might help to understand better the nature of the measurement errors when using PCA and EFA composites.

Assume that the *underlying* model to be estimated is the following Cobb-Douglas production function:

$$\ln y = \beta_0 + \beta_1 \ln x_1 + \beta_2 \ln x_2 + \beta_3 \ln x_3 \tag{40}$$

Note that the β coefficients capture the effect of each input on y. From a theoretical point of view, this effect does not rely on how both inputs are aggregated using statistical techniques. An equivalent way to write the model above using a single *theoretical* input composite X is:

$$\ln y = \beta_0 + \beta \ln X, \quad X = \prod_{n=1}^{3} x_n^{\beta_n/\beta} \tag{41}$$

where $\beta = \beta_1 + \beta_2 + \beta_3$ measures economies of scale, i.e., the overall effect of all inputs on firm's production. The latter specification simply indicates that inputs should be weighted using their relative output-elasticity (that in turns depends on the relative productivity) in order to estimate properly the scale elasticity β. In other words, the *empirical* composite should aggregate x_1, x_2 and x_3 taking *somehow* into account the dependent variable. This is precisely the aim of the *supervised* methods for reducing the dimension of the data that are outlined later on.

[41] A comprehensive discussion about separability and the theoretical implications of different types of separability (e.g., strong vs. weak) can be found in Chambers (1988).

In contrast, PCA and EFA are *unsupervised* because they only use information on how x_1, x_2 and x_3 are statistically distributed, how large their variances are, or whether they are highly correlated. Therefore, predictions using simple statistical techniques might be biased because relevant predictive variables can be underweighted, while irrelevant factors can be overweighed. This type of error might explain the fact that clear relationships are not often obtained in many studies using PCA and EFA composites.

In an attempt to better understand why these composites are not able to capture the underlying output-elasticities of each input, let us now assume that we are using two PCA composites, X_1 and X_2, which are linear combinations of the three original inputs:

$$\begin{aligned} \ln X_1 &= \theta_{11} \ln x_1 + \theta_{12} \ln x_2 + \theta_{13} \ln x_3 \\ \ln X_2 &= \theta_{21} \ln x_1 + \theta_{22} \ln x_2 + \theta_{23} \ln x_3 \end{aligned} \tag{42}$$

Thus, in this case the estimated production function is:

$$\ln y = \gamma_0 + \gamma_1 \ln X_1 + \gamma_2 \ln X_2 \tag{43}$$

We next compare the estimated parameters and the underlying parameters that result from plugging the input composites in (42) into (43). After ordering terms this yields:

$$\ln y = \gamma_0 + \left(\gamma_1\theta_{11} + \gamma_2\theta_{21}\right)\ln x_1 + \left(\gamma_1\theta_{12} + \gamma_2\theta_{22}\right)\ln x_2 + \left(\gamma_1\theta_{13} + \gamma_2\theta_{23}\right)\ln x_3, \tag{44}$$

or

$$\ln y = \gamma_0 + \delta_1 \ln x_1 + \delta_2 \ln x_2 + \delta_3 \ln x_3 \tag{45}$$

where δ_1, δ_2 and δ_3 can be interpreted as the implicit PCA parameters of each input. From (45) it is straightforward to see that each implicit parameter shares components with other implicit parameters, and hence the model based on PCA composites can be seen as a restricted least squares estimator.[42] In this sense, Fomby et al. (1978) show that the principal component estimator is the restricted least squares estimator with the smallest variance of those possessing the same number of restrictions. The small variance is a virtue, but as shown by Greene (2002) it is a biased estimator.

4.1.2. Supervised dimension reduction techniques

The variable reduction technique is labelled as *supervised* or sufficient when the relationship between the variable to be predicted and the vector of explanatory variables to be aggregated is used as information. See Cook (2007) and Adragni and Cook (2009) for a formal definition, and Bura and Yang (2011) for an overview of sufficient dimension reduction in regression. Li (1991) introduced the first method for sufficient dimension reduction, i.e., sliced inverse regression (SIR), and since then various types of inverse regressions have been proposed.

The inverse regression methods have been intensively applied in fields such as biology, genome sequence modelling, and pattern recognition involving images or speech. However, the potential of sufficient dimension reduction methods for reducing datasets has barely been explored

[42] This is trivial in the case of a unique composite. In this case the implicit parameter of each input can be written in terms of other implicit parameters, in particular, as $\delta_n = \delta_1 \cdot (\theta_{11}/\theta_{1N})$, $N \neq 1$.

in economics. An exception is Naik et al. (2000) that use the SIR techniques to aggregate marketing data, and Orea et al. (2015) that is the first attempt to apply supervised methods to production economics using a SFA model.

The popularity of inverse regression methods in other fields is due to the fact that most of them are computationally simple. We next describe briefly the procedure of the original sliced inverse regression method. Simplifying the notation in Li (1991), the model to be estimated can be written as:

$$\ln y = \beta_0 + \beta_1 \ln f(\theta X) + \varepsilon = \beta_0 + \beta_1 \ln f\left(\Sigma_{n=1}^{n=2} \theta_n x_n\right) + \varepsilon \,, \tag{46}$$

where firms' output y is the response variable, $X=(x_1,x_2)$ now is a 2-dimensional vector of inputs that are to be aggregated, $\theta=(\theta_1,\theta_2)$ is a vector of unknown coefficients, and ε is a random term which is assumed to be independent of the inputs levels. This method makes no assumption about the distribution of the error term. This makes it appealing for SFA applications where the error term includes noise and inefficiency random terms. In this formulation, the response variable is related to the *2*-dimensional regressor X only through the reduced 1-dimensional variable θX. If there were more inputs, there could be more θ vectors. In this case, if N is the dimension of the vector of inputs, the dimension of θ should be $H<N$, that is $\theta=(\theta_1,\ldots,\theta_H)$ where in turn each θ_i is a N-dimensional vector.

SIR and other sufficient dimension reduction methods are developed to find the space generated by the unknown θ vector. This space should be estimated from the data and is based on the spectral decomposition of a kernel matrix K that belongs to the central subspace (i.e., the smallest dimension-reduction subspace that provides the greatest dimension reduction in the predictor vector). Most sufficient dimension reduction methods are based on the same logic but use different kernel matrices. Once a sample version of K is obtained, the corresponding eigenvectors are used to estimate the central subspace. To this aim, Li (1991) proposed to reverse the conventional viewpoint in which y is regressed on X, and showed that a principal component analysis on a nonparametric estimate of $E(X|y)$ can be used to estimate K. The approach relies on partitioning the whole dataset into several slices according to the y-values. Thus, the dependent variable is only used to form slices while the standard PCA does not use any information from y.

The above procedure provides a rather crude estimate of $E(X|y)$. If the response variable is a continuous variable, its transformation to a categorical variable might imply ignoring much information about the inverse relationship between the set of inputs and the output variable. Parametric inverse regression (PIR) is another first-moment based method for effective dimension reduction that aims to estimate the central subspace using least squares. This parametric version of inverse regression regresses each (standardized) input variable on a set of arbitrary functions of y. The fitted values of $E(X|y)$ are then used as an estimate of K.

4.1.3. The choice between supervised and unsupervised methods

Regarding the choice between supervised and unsupervised variable dimension reduction methods, it is apparent from the above discussion that we will always get better results using a supervised method than an unsupervised one. In theory this is true, but not in practice. Note that the *supervised* methods need to control the relationship between the variable to be predicted and the vector of explanatory variables when they proceed with the aggregation. In practice, this relies

on a principal component analysis of a (non)parametric estimate of $E(X|y)$.[43] In this sense, Adragni and Cook (2009) pointed out that some of the best moment-based methods turned out to be rather inefficient in relatively simple settings. Thus, in any particular efficiency analysis it could occur that $E(X|y)$ is too poorly estimated meaning that an *unsupervised* method might yield better results. In order to minimize biases associated to inaccurate inverse regressions, we thus suggest using more recent, but more complex *supervised* methods.[44] Even then, a limitation of these variable dimension reduction methods in efficiency analysis is that their sample property results have been obtained with normally distributed error components. A field of future research is the analysis of their properties in SFA models with asymmetric error terms.

4.1.4. The choice of the number of composites

A final issue that should be examined in this subsection is determining the number of composites to retain. To choose the number of composites we propose using conventional model selection tests that balance the lack of fit (too few composites) and overfitting (too many composites). The use of model selection tests is usually restricted to cases where economic theory provides no guidance on selecting the appropriate model, and the alternative models are not nested, as in the present approach.

Again we have to make a choice as there are many model selection tests. In addition to the traditional statistics summarized in the next subsection, we should mention the so-called sequential tests (ST) method introduced in the literature on sufficient dimension reduction (see Li, 1991; and Bura and Cook, 2001). Unlike conventional model selection tests, the ST approach tries to determine the dimension of the central subspace independently from the estimation of the production function and, hence, does not take into account the prediction of y. One limitation of the ST tests is that most of them have been developed for data visualization and they tend to retain just one or two composites in most regression problems. Another limitation is that the significance level at each step in a sequential test procedure does not determine the significance level of the entire procedure, and that the retained dimension depends on the choice of significance level. Moreover, Zhu et al. (2006) carried out several simulations to examine the relative performance of both BIC-type methods and ST tests, and found that BIC-type methods are clearly better than ST methods.

4.2. Variable selection

In biology, industrial engineering and other non-economic fields where hundreds or even thousands of noisy and/or correlated explanatory variables are available to predict a response variable, variable selection procedures are well recognized. Indeed, the elimination of variables is highly desirable in these fields as they are mainly interested in predicting a response (dependent)

[43] Sliced average variance estimate (SAVE) and other less popular *second*-moment based methods can be considered. These methods may also capture additional information that is orthogonal to the first moment of $X|Y$ (Cook and Weisberg, 1991). However, as they look at second moment information, they can only detect nonlinear patterns, and hence they often give poor predictive composites.

[44] For instance, whereas Xia et al. (2002) and Bura (2003) proposed semiparametric techniques to estimate the inverse mean function, $E(X|Y)$, Cook and Ni (2005) developed a family of dimension-reduction methods by minimizing quadratic discrepancy functions and derived the optimal member of this family, the inverse regression estimator (IRE).

variable and one cost of overfitting (i.e., estimating a more complex model than it needs to be) is the increased variability of the estimators of the regression coefficients.

In economics, we should at least add three additional reasons to proceed with the elimination of variables. First, the 'dimensionality' issue becomes acute when flexible functional forms are estimated as the dimensionality increases more rapidly when interactions are considered or the semi-parametric or non-parametric techniques require a manageable number of explanatory variables to be implemented. Second, for interpretation reasons: the identification of relevant variables based on economic theory or expertise' knowledge may or may not be correct if the model is overfitted. Finally, it is always preferable to build a parsimonious model for easier data collection. This is especially relevant in efficiency analyses in regulated industries, such as the electricity distribution sector, where the regulators need to collect costly data on a large set of variables in order to control for many geographical, climatic or network characteristics of the utilities sector that affect production costs, but which go unobserved.

4.2.1. Standard variable selection techniques

This subsection does not intend to explain the statistical basics of techniques used for variable selection or explore for that matter entire approaches, but instead seeks to provide researchers and practitioners with a description of representative methods recently developed for selecting variables. Many different procedures and criteria for selecting the best regression model have been suggested. See Mittelhammer et al. (2000) for a general and critical analysis of the variable selection problem and model choice in applied econometrics and Anzanelo and Fogliatto (2014) in industrial-type applications.

The basic selection techniques are backward, forward and stepwise procedures. In the *backward procedure*, the researcher starts with a full model containing all variables and them goes on to use a significance test based on the F statistic to sequentially remove variables, thereby selecting the "reduced" model. The process is repeated until the remaining variables have an F value larger than a defined threshold. The resulting variables are then used in the prediction model. This procedure, and other more comprehensive methods, becomes impractical with hundreds of variables. The *forward procedure* is similarly guided by a significance test, but now variables are introduced one-by-one into the model according to their F or p-values. The *stepwise procedure* modifies the forward selection procedure in that each time a new variable is added to the model, the significance of each of the variables already in the model is re-examined. The stepwise regression procedure continues until no more variables can be added or removed.

These simple procedures may lead to interpretable models. However, as Mittelhammer et al. (2000) note, the results can be erratic as any single test used at some stage in the above-mentioned procedures is not indicative of the operating characteristics of the joint test represented by the intersection of all the individual tests used. That is, because the subset selection is a discrete process, small changes in the data can lead to very different statistical models, and the sampling properties of these processes are virtually unknown.

In addition to the F and p-value, other criteria have been used to evaluate the impact of adding/removing variables in regression models. These include the Akaike's criterion (AIC) (Akaike, 1974), the Schwarz's Bayesian criterion (SBC) (Schwarz, 1978), and some of their

variants, which penalize the model as new explanatory variables are added.[45] The general form of most information criteria can be written as follows:

$$-2\ln LF + Penalty\,, \tag{47}$$

where the first term is twice the negative logarithm of the maximum likelihood which decreases when the number of variables (complexity) increases. For more details about these criteria and the associated penalty functions, see Fonseca and Cardoso (2007). The penalty term penalizes very complex models, and increases with the number of parameters of the model. Thus, these criteria involve minimizing an index that balances the lack of fit (too few variables) and overfitting (too many variables). Models with lower values of (47) are generally preferred.

The above model selection procedures have been criticized because the deterministic nature of these criteria means that no information is provided as to ''how much'' better the chosen model is (i.e., they do not allow probabilistic statements to be made regarding model selection). For this reason, Vuong's (1989) model selection framework is often advocated to select the most adequate model. This test is designed to test the null hypothesis that two competing models fit the data equally well versus the alternative that one model fits them better. Initially, the Vuong test can be used to choose between two models whether they are nested or not. But the form of the statistic and the distribution theory are different. In particular, in the case of nested models, we have a nonstandard (mixture of chi-squared distributions) distribution.[46] For this reason it is typically not used for nested models.

It should be pointed out that variable selection and dimensionality reduction also carry a long tradition in the DEA literature. There are methods that study correlation among variables, with the goal of choosing a set that do not represent largely associated values. However, these approaches may yield unreliable results because the removal of even highly correlated variables can still have a large effect on the DEA results− Nunamaker (1985). For instance, Lewin et al. (1982) and Jenkins and Anderson (2003) apply regression and multivariate analysis to select and reduce the number of variables in the DEA model, respectively. In the former case, when evaluating the administrative efficiency of courts, the authors resort to two types of regressions to determine the appropriate model specification capturing the relationship between outputs, inputs and contextual−non discretionary−variables (such as *per capita* income or percentage of white population). They run two linear regressions imposing constant returns to scale, and a log-log specification allowing for variable returns to scale. In both cases they study the explanatory power of the variables using a step-wise method by which variables are included sequentially. The latter regression yields a better goodness-of-fit, and based on these results, they select the specific inputs, outputs and non-discretionary variables to be included in the DEA model. It is interesting to note that Dyson et al. (2001), when studying several pitfalls and protocols in DEA, call for exercising caution when simply dropping some variables based on their high correlation (e.g., inputs) since efficiency scores can change significantly, this result being illustrated with a simple example. As

[45] Other proposals on model selection include the risk inflation criterion (Foster and George, 1994), the generalized information criterion (Konishi and Kitagawa, 1996), a Bayesian measure using the deviance information criterion (Spiegelhalter et al., 2002; Gelman et al., 2004), model evaluation using the absolute prediction error (Tian et al., 2007), tuning parameter selection in a penalization method (Wang et al., 2007; Fan and Tang, 2013), the 'parametricness' index (Liu and Yang, 2011), and many extensions of AIC and BIC—see, for example, Bozdogan (1987, 2000), Bogdan et al. (2004), Chen and Chen (2008) and Zak-Szatkowska and Bogdan (2011)).

[46] We thank Peter Schmidt for relevant comments on this matter.

the sequential regression method suggested by Lewin et al. (1982) is influenced by the collinearity between regressors, it is prone to the previous selection problem.

To precisely address the arbitrariness and problems related to discarding variables based on their high correlations with those ultimately retained in the analysis, Jenkins and Anderson (2003) propose a multivariate analysis to reduce the dimensionality of variables. After reviewing previous proposals based on multivariate analysis, such as canonical correlation (Sengupta, 1990; Friedman and Sinuany-Sterns, 1997), discriminant analysis (Sinuany-Sterns and Friedman, 1998), and the already mentioned principal components analysis (Ueda and Hoshiai, 1997; Adler and Golany, 2001, 2002), these authors propose a multivariate method to identify which variables can be discarded with least loss of information. The method is based on the variance of the input or output about its mean value, for if its value is constant, then it plays no part in distinguishing one DMU from another. On the contrary, a large variation indicated an important effect. Their method uses partial correlation as a measure of information contained in each variable with respect to their counterparts. If a variable is perfectly correlated with another, the remaining variance when one of the two is removed is zero. Authors normalize the variables to have zero mean and a variance of 1, so as to scale the total combined variance to the number of inputs and outputs, n and m, and calculate the conditional variance dropping input variables sequentially, starting with those presenting the highest partial correlation coefficients and following in decreasing order (the method can be applied to outputs). Comparing the results obtained with their method using the databases of several published studies, confirms the worries expressed by Dyson et al. (2001) as there are large variations in the computed efficiencies.

A second strand of literature examines whether removing or adding variables in a sequential way results in significant changes in the DEA efficiency distributions −Norman and Stoker (1991). In this approach, variables are to be discarded or included according to a selection process that assesses their statistical significance. In this vein Kittelsen (1993) surveys several tests to establish the significance of change in efficiency results when sequentially removing or adding variables. He shows that the usual tests (F, Kolmogorov-Smirnov, t-ratio, etc.) to determine if the subsequent efficiency distributions remain the same or change after removing or adding variables are invalid because they assume that the scores are independently and identically distributed. This is not the case with the sequential method because the individual DEA efficiencies are nested, with those corresponding to the augmented model including more variables (restrictions) presenting larger scores−a property deriving from linear optimization theory, and resulting also in more efficient firms. Applying the above tests to model selection results in biased results−a problem aggravated by small samples, but it is possible to control for the relative bias and use these tests, by examining the sample size, dimensionality and efficiency distributional forms. Kittelsen then proposes a stepwise sequential process (algorithm), starting with a model including all the variables, and removing those that exhibit lower significance according to the above tests, until a satisfactory model specification with all significant variables is achieved. He exemplifies the method for a database of Norwegian electric distribution utilities.

This is a valuable exercise that nevertheless should be revisited with the use of tests that take into account the nested nature and distributional forms of the efficiency scores as proposed by Pastor et al. (2002). These authors define a new ‘efficiency contribution measure’ (ECM, representing the marginal impact on efficiency of a variable), that compares the efficiency scores of two radial DEA models differing in one output or input variable (termed candidate). Then, based on this ECM, at a second stage a statistical test is developed that allows an evaluation of the

significance of the observed efficiency contribution of the differing variable (i.e., the effects above a level of tolerance or threshold). This test provides useful insights for the purpose of deciding whether to incorporate or delete a variable into/from a given DEA model, on the basis of the information supplied by the data. Two procedures for progressive selection of variables are designed by sequentially applying the test: a forward selection and a backward elimination.

This method for selecting the variables to be included in the model is reworked by Wagner and Shimshak (2007), who improve these procedures by formalizing a stepwise method which shows how managers can benefit from it in a structured decision making process. As in the parametric framework, their method suggests some simple rules for sequentially removing variables (backward elimination of those that have the least influence from amongst the set of efficient DMUs defining the frontier) or for adding variables (forward selection of the most influential ones), one at a time. The backward elimination algorithm proposed by these authors can be summarized as follows:

1) Calculate the efficiency scores for the preferred complete DEA model E^* including all N inputs and M outputs.
2) Calculate the efficiency scores of the alternative N-1 and M-1 DEA models, $E^*_{x_n}$ and $E^*_{y_m}$, where a single n-th input or m-th output are removed at a time in each run.
3) Remove the n-th input or m-*th* output for which the average difference between the complete and partial scores is the smallest: $\min (E_{x_1} = \sum_{i=1}^{I} (E^* - E^*_{x_1}) / I, ..., E^d_N = \sum_{i=1}^{I} (E^* - E^*_N) / I,$ $E_{y_1} = \sum_{i=1}^{I} (E^* - E^*_{y_1}) / I, ..., E_{y_M} = \sum_{i=1}^{I} (E^* - E^*_{y_M}) / I)$.
4) Departing from the efficiency scores associated to the model without the dropped input or output, repeat the process until, in principle, only one input and output remain in the model.

As it is not desirable to remain with one single input and output (the "core" model), the authors propose stopping the removal of variables (or addition) when the average difference of efficiency scores, or any individual difference, exceeds a maximum level. They exemplify their method with several datasets previously used in this literature and compare their results with those previously attained; e.g., the academics department data used by Sinuany-Stern et al. (1994) and Jenkins and Anderson (2003) above.

4.2.2. Model selection with high dimensional datasets

We now focus on variable selection procedures involving hundreds or thousands of explanatory variables, i.e., model selection with high dimensional datasets. The traditional methods face significant challenges when the number of variables is comparable to or larger than the sample size. These challenges include how to make the estimated models interpretable, in our case, from an economic perspective.

An alternative approach to proceed with variable selection with high dimensional datasets is penalized least squares, which is a method for simultaneous estimation and variable selection, (see Fan and Li, 2006, and Fan and Lv, 2010 for overviews). All of them try to minimize a so-called penalized least squares function that tends to produce some coefficients that are exactly zero. As this outcome is equivalent to a reduction in candidate explanatory variables from the

model, LASSO and other penalized least squares estimators help in getting more interpretable models.

As pointed out by Fan and Lv (2010), what makes high dimensional statistical inference possible is the assumption that the *underlying* (distance) function does have less variables than the dataset. In such cases, the *d*-dimensional regression *parameters* are assumed to be sparse with many components being zero, where nonzero components indicate the important variables. With sparsity, variable selection can improve the estimation accuracy by effectively identifying the subset of important predictors and can enhance the model interpretability with parsimonious representation.

Many variable selection criteria or procedures are closely related to minimizing the following penalized least squares (PLS):

$$\frac{1}{2I}\sum_{i=1}^{I}\left(y_i - x_i\beta\right)^2 + \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|), \tag{48}$$

where d is the dimension of x_i, and $p_\lambda(\cdot)$ is a penalty function indexed by the regularization parameter $\lambda \geq 0$, controlling for model complexity. The dependence of the penalty function on j is very convenient in production and cost analyses as it allows us to keep certain important explanatory variables in the model (for instance, key inputs in a production function, or the output and input prices variables in a cost function) thus choosing not to penalize their coefficients.

The form of the penalty function determines the general behaviour of the estimator. With the entropy or L_0-penalty, namely $p_{\lambda_j}(|\beta_j|) = \frac{1}{2}\lambda^2 \mathbf{1}_I(|\beta_j| \neq 0)$, the PLS in (48) becomes

$$\frac{1}{2I}\sum_{i=1}^{I}\left(y_i - x_i\beta\right)^2 + \sum_{j=1}^{d}\lambda^2 |M|, \tag{49}$$

where $|M|$ is the size of the candidate model. In this formulation, among models with the same number of variables, the selected model is the one with the minimum residual sum of squares. Many popular variable selection criteria have been shown asymptotically equivalent to the PLS (49) with appropriate values of λ (see, for instance, Fan and Lv, 2010).

Many researchers have been working on minimizing the PLS (49) with *Lp*-penalty for some $p>0$. While the L_2-penalty results in a ridge regression estimator, the *Lp*-penalty with $0<p<2$ yields bridge regression (Frank and Friedman, 1993). With the L_1-penalty specifically, the PLS estimator is called LASSO in Tibshirani (1996). When $p \leq 1$, the PLS automatically performs variable selection by removing predictors with very small estimated coefficients. None of the above *Lp*-penalties satisfy all three desirable conditions simultaneously: sparsity, unbiasedness and continuity. The LASSO estimator satisfies the sparsity condition as it should automatically set small estimated coefficients to zero in order to accomplish variable selection. However, it is a biased estimator, especially when the underlying coefficient of dropped variables is large.[47]

The PLS approach can also be easily extended to the likelihood framework. Define a penalized likelihood function as:

[47] It should be noted that there are many other penalty functions satisfying the aforementioned three conditions. For instance, the smoothly clipped absolute deviation (SCAD) penalty function introduced by Fan (1997) and Fan and Li (2001).

$$Q(\beta) = \frac{1}{I}\sum_{i=1}^{I} \ln LF\left(y_i, x_i; \beta\right) - \sum_{j=1}^{d} p_{\lambda_j}(|\beta_j|) \tag{50}$$

Maximizing the penalized likelihood results in a penalized likelihood estimator. For certain penalties the selected model based on the penalized likelihood satisfies β_j=0 for certain β_j's. Therefore, parameter estimation is performed at the same time as the model selection. As the likelihood framework is the most used framework in the SFA literature, we believe that this is a promising area of research for the near future when we will progressively be able to collect large and larger data sets to carry out efficiency analyses.

4.2.3. The choice between variable dimension reduction and variable selection

We conclude this section with a practical discussion concerning the choice between variable dimension reduction and variable selection in efficiency analyses. Indeed, as *all* explanatory variables in variable dimension reduction have only an indirect effect on the dependent variable, this empirical strategy can be specially advocated when: (i) we are trying to control for a holistic phenomenon formed by a large number of factors with complex interactions, and (ii) it is difficult to either formulate hypotheses associated to these variables or impose restrictions derived from production theory on the technology.

In other words, variable dimension reduction is appealing when the main issue is the overall effect of a wide-ranging phenomenon, and not the partial effect of its components. In this sense, environmental variables are good candidates for aggregation using some of the techniques outlined above. On the other hand, this approach is probably more suitable in DEA applications where researchers' main interest is in measuring firms' inefficiency and not in disentangling *specific* technological characteristics such as economies of scale, scope, substitution between inputs/outputs, etc.

Finally, it is worth mentioning that many electricity, gas, and water distribution firms are currently regulated using incentive-based regulation regimes that rely on frontier techniques. Regulators often use very simple DEA and SFA models that are easy to understand by the regulated firms in order to reduce the risk of litigation and prevent judicial trials. As composite variables normally lack measurement units and a clear economic interpretation, they often avoid using variable dimension reduction and prefer using alternative empirical strategies such as the variable selection approach outlined above.

4.3. The choice of functional form

Consistent with the axioms of the technology, 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 Constant Elasticity of Substitution (CES), as well as their associated dual cost or (by extension) profit functions, place significant restrictions on technological and economic behaviour relations. For example, in production analysis they restrict all elasticities of substitution to be equal to one for all inputs, or to the same value across them, respectively; while for cost minimization, the linear or log-linear specifications imply that inputs

demands, or the share of each input in costs, are independent of the output level, evidencing that the corresponding technology is restricted to the homothetic case with ray-linear expansion paths.[48]

While these characteristics are quite restrictive, these functions meet the regularity conditions that make them suitable representations of technology through distance or production functions: namely, they are "well-behaved" and satisfy all desirable neoclassical properties, particularly quasi-concavity, which ensures that the associated production possibility sets are convex, Madden (1986); and that, for analytical convenience, they are continuous and twice differentiable.[49] 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 and economic behavioural models (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 (Gallant, 1981, 1984). One advantage of the latter proposal is that it provides a global rather than a local approximation (e.g., quadratic specifications fail to satisfy the regularity conditions over the entire range of sample observations); but since its econometric estimation and parameter interpretation prove more demanding, they are by far less popular in empirical research. For these reasons in this section we restrict our discussion to the most widespread and customarily chosen second order functional forms. In any case, following Griffin et al. (1987) the choice of functional form based on alternative criteria should be treated explicitly in empirical research; e.g., of paramount importance, from a production analysis perspective, is the ease with which to impose the homogeneity and translation properties characterizing the radial, directional or generalized distance functions, as previously discussed in the Section 3.2 and exemplified below.[50]

Let us consider that the true, twice continuously differentiable, distance (production), cost, profit or profitability function is represented by the functional form $f(x)$. Then, a second order approximation of this function can be provided by another function $g(x)$ at an arbitrary point x^* in the domain of definition of $f(x)$ and $g(x)$. This implies that given the second order expansion of the true function around x^*,

[48] The limitations of the Cobb-Douglas functions when testing the neoclassical theory of the firm have been extensively criticized (Samuelson, 1979), and have constituted the basis for newer, less restrictive proposals such as the CES and other functional forms (Zellner and Revankar, 1969). Much effort was devoted to relaxing some of the constraints associated to these classical functions; e.g., Sato (1977) introduced their general non-homothetic versions. However, a qualitative step forward took place with the introduction and popularization of the notion of flexible functional forms, whose parametrization allows the imposition and testing of relevant properties relating to the technology and economic behaviour. For a thoughtful discussion of alternative functional forms see Thompson (1988).

[49] The regularity and differentiability conditions of the production function passes on to the more general distance functions representation–see Blackorby and Diewert (1979).

[50] For a thoughtful discussion of alternative functional forms see Thompson (1988).

$$f\left(x^{*}+\Delta x\right)\approx f(x^{*})+\sum_{n=1}^{N}\frac{\partial f(x^{*})}{\partial x_{n}}\Delta x_{i}+\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\frac{\partial^{2} f(x^{*})}{\partial x_{n}\partial x_{n'}}\Delta x_{n}\Delta x_{n'}, \tag{51}$$

and for infinitesimal variations, $f\left(x^{*}+\Delta x\right)$ can be closely approximated in terms of the level of f at x^{*}, $f\left(x^{*}\right)$, and its first and second derivatives at x^{*}: $\left(\frac{\partial f(x^{*})}{\partial x_{n}},\frac{\partial^{2} f(x^{*})}{\partial x_{n}\partial x_{n'}}\right)$. In this case, $g\left(x\right)$ provides a second order (differential) approximation of $f\left(x\right)$ at x^{*} if and only if the following $1+N+N^{2}$ equations are satisfied: $g\left(x^{*}\right)=f\left(x^{*}\right)$, $\frac{\partial g(x^{*})}{\partial x_{n}}=\frac{\partial f(x^{*})}{\partial x_{n}}$, and $\frac{\partial^{2} g(x^{*})}{\partial x_{n}\partial x_{n'}}=\frac{\partial^{2} f(x^{*})}{\partial x_{n}\partial x_{n'}}$.

Consequently, a general second order flexible functional form $g\left(x\right)$ must have enough free parameters. Since by using Young's theorem from calculus the cross-partial derivatives of the true and approximating functions are symmetric: $\partial^{2} g(x^{*})/\partial x_{n}\partial x_{n'}=\partial^{2} g(x^{*})/\partial x_{n'}\partial x_{n}$, then the number of parameters reduces to $1+N+N(N+1)/2$. Additionally, if $f\left(x\right)$ and $g\left(x\right)$ are linear homogeneous in x as in the case of constant returns to scale in distance or production functions, then (using Euler's theorem) implies additional restrictions on the $1+N$ equations: $g\left(x^{*}\right)=\sum_{i=1}^{N}\left(\partial g\left(x^{*}\right)/\partial x_{n}\right)x_{n}^{*}$, $0=\sum_{n=1}^{N}\left(\partial^{2} g\left(x^{*}\right)/\partial x_{n}\partial x_{n'}\right)x_{n'}^{*}$.

While the flexibility of the functional forms allows a more precise representation of the production technology and economic behaviour, they are prone to some drawbacks. The fact that the number of parameters to be estimated increases exponentially with the number of variables included in the functional form (e.g., quantities of inputs and outputs, prices, etc.), and the order of the Taylor series approximation (second, third, etc.), empirical research is *de facto* restricted to the quadratic approximation, which nevertheless requires large samples to meet enough degrees of freedom for statistical hypotheses testing. 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, and runs into problems of its own (as discussed in the previous section).

Finally, as earlier remarked, the properties of flexible functional forms become relevant as they ultimately determine whether they are globally well-behaved in the presence of large data variability. However, how to test those global properties and impose regularity conditions globally remains unclear. For instance, Lau (1986) proved that flexibility is incompatible with global regularity if both concavity and monotonicity are imposed using standard econometric techniques. That is, 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. In the end, it is hard to determine the restrictions that these functional forms impose on *a priori* grounds, and the consequences that they have on econometric estimates (see, for instance, Fuss et al., 1978; and Saunders, 2008).

It should be pointed out, however, that it is possible to maintain *local* flexibility of these flexible functions using Bayesian techniques. In this sense, and focusing on the cost function,

Terrell (1996) and Griffiths et al. (2000) present a Bayesian approach which imposes monotonicity and concavity properties only over the set of prices where inferences will be drawn. O'Donnell and Coelli (2005) later on proposed using a Bayesian approach to impose the monotonicity, quasi-convexity and convexity constraints implied by economic theory on the parameters of a Translog output distance function. The local approach maintains the flexibility property of a functional form if the regularity conditions are imposed at each observed regressor value. Although the procedure becomes numerically difficult for large sample sizes and/or complicated constraints when compared with the global approach, this approach generally increases the fit of the model to the data and leads to better forecasts than global regularity.

Despite these caveats, flexible functional forms are useful and have become standard in empirical studies given the restrictions imposed by the CD and CES formulations, which should not be favoured in empirical microeconometric research given their limitations. To exemplify their capabilities when testing functional properties such as returns to scale, homotheticity, etc., we show two representative specifications corresponding to the primal and dual approaches. The first one makes use of the *Quadratic* formulation to specify the directional distance function, and the second one corresponds to the *Translog* cost function.

4.3.1. Quadratic directional distance functions

As for the output 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, output, or hyperbolic distance functions can be easily imposed on the Translog specification. For convenience we set the directional vector $(-g_x, g_y)$ to (-1, 1). The interpretation of the resulting distance function values yields net optimal output and input vectors. This can be easily seen taking the difference between the optimal and observed output and input vectors: i.e., $y + D_T(x, y; -1, 1) \cdot 1_M - y$ and $x + D_T(x, y; -1, 1) \cdot 1_N - x$, and the amount in which outputs can be increased and inputs reduced corresponds to the value of the distance function. Additionally, from the perspective of aggregation, adding all individual inefficiencies yields the whole industry savings in terms of inputs and additional gains in terms of larger output (as a proxy of cost savings and revenue increases can be obtained resorting to market prices). Turning now to the parametric specification, it corresponds to

$$D_T(x, y; -1, 1) = \alpha_0 + \sum_{n=1}^{N} \alpha_n x_n + \sum_{m=1}^{M} \chi_m y_m + 1/2 \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_{nn'} x_n x_{n'} + 1/2 \sum_{m=1}^{M} \sum_{m'=1}^{M} \chi_{mm'} y_i y_j + 1/2 \sum_{n=1}^{N} \sum_{m=1}^{M} \delta_{nm} x_n y_m \text{ , (52)}$$

with symmetry between cross parameters: $\alpha_{nn'} = \alpha_{n'n}, \chi_{mm'} = \chi_{m'm}, \forall i, j$. For the translation property introduced in the second section to hold: $D_T(x - g_x\lambda, y + g_y\lambda; g_x, g_y) = D_T(x, y; g_x, g_y) - \lambda, \lambda \in \mathbb{R}$, the required parameter restrictions are: $\sum_{m=1}^{M} \chi_m g_{ym} - \sum_{n=1}^{N} \alpha_n g_{yn} = -1$, $\sum_{n}^{N} \alpha_{nn'} g_{xn} = 0, n = 1, ..., N$, $\sum_{n}^{N} \alpha_{nn'} g_{xn'} = 0, n' = 1, ..., N$, $\sum_{m}^{M} \chi_{mm'} g_{ym} = 0, m = 1, ..., M$, $\sum_{m}^{M} \chi_{mm'} g_{ym_i} = 0, m' = 1, ..., M$ and $\sum_{n}^{N} \delta_{nm} g_{xn} = \sum_{m=1}^{M} \delta_{nm} g_{ym} = 0, n = 1, ..., N, m = 1, ..., M$.

4.3.2. Translog cost function

As for the Translog cost function, the specification corresponds to:

$$\ln C(y,w) = \alpha_0 + \sum_{m=1}^{M} \alpha_i y_i + 1/2 \sum_{m=1}^{M} \sum_{m'=1}^{M} \alpha_{mm'} \ln y_m \ln y_{m'} + \sum_{n=1}^{N} \chi_n p_n + 1/2 \sum_{n=1}^{N} \sum_{n'=1}^{N} \chi_{nn'} \ln p_n \ln p_{n'} + 1/2 \sum_{m=1}^{M} \sum_{n=1}^{N} \delta_{mn} y_m p_n, \quad (53)$$

with the corresponding symmetry restrictions $\alpha_{mm'} = \alpha_{m'm}, \chi_{nn'} = \chi_{n'n}$. To be consistent with theory the following restrictions on the parameters, securing homogeneity of degree 1 in prices, are necessary: $\sum_{n=1}^{N} \chi_n = 1$, $\sum_{n=1}^{M} \delta_{nm} = 0, m = 1,\ldots,M$, $\sum_{n=1}^{N} \chi_{ij} = \sum_{n'=1}^{N} \chi_{ji} = \sum_{n=1}^{N} \sum_{n'=1}^{N} \chi_{nn'} = 0$. Additional technological properties related to homotheticity (i.e., separability between outputs and factor prices), and homogeneity (i.e., whether cost elasticity is constant with respect to output), could also be imposed, even if the best strategy is not to impose them, but test them empirically after estimation through hypotheses testing. In that case, for homotheticity, it must be checked if $\sum_{m=1}^{M} \alpha_m = 1$, while homogeneity requires both $\sum_{m=1}^{M} \alpha_m = 1$ and $\sum_{m=1}^{M} \alpha_{mm'} = 0$. Further parameter restrictions regarding outputs and inputs elasticities of transformation and substitution, respectively, could be tested.

While additional flexible functional forms for profit could be presented,[51] these two examples show the attractiveness of second order approximations for discerning technological properties through primal or dual specifications, and the need to check the properties of the functional forms, so as to choose the most appropriate form in empirical analysis. For instance, from an economic perspective, it is necessary to determine the nature of returns to scale in the light of the theoretical results (i.e., variable for cost minimization and revenue and profit maximization, and constant for profitability maximization). Here, homotheticity is customarily assumed throughout, even if it is a key property when interpreting and decomposing economic efficiency according to technical and allocative criteria and as such, should not be imposed on the specification, but rather be tested through the relevant parameter hypotheses thus allowing the data to reflect it (Aparicio and Zofío, 2017).

5. Controlling for observed environmental conditions

Obtaining reliable measures of firms' efficiency requires controlling for the different environmental conditions under which each firm operates. This is especially acute in regulated markets based on benchmarking because of the financial implications that this analysis has over the regulated firms and their effect on the whole industry. The concern about the inclusion of environmental variables (also called *contextual*, non-discretionary or *z*-variables) has generated the development of several models either using parametric, nonparametric or semi-parametric techniques. Although we do not pretend to provide a complete survey of the alternatives for including *z*-variables, given the wide range of models that have been developed, here we only mention the methods most frequently applied. For a more detailed review of this topic in SFA, see Parmeter and Kumbhakar (2014). A brief summary of this issue in the non-parametric literature can be found in Johnson and Kuosmanen (2012).

[51] Coyle (2010; ch. 5) presents the formulations for the dual quadratic, generalized Leontief and translog profit functions.

5.1. DEA estimation of the effects of contextual variables

The inclusion of environmental variables in DEA has been done in one or two stages.[52] The one-stage DEA approach (hereafter 1–DEA) is to augment the model by treating the z–variables as inputs or outputs that contribute to defining the frontier. For instance, the directional distance function *with* environmental variables would be:

$$D_V(y,x,z;-g_x,g_y)=\max\left\{\beta:(x-\beta g_x,y+\beta g_y)\in V(z)\right\}. \tag{54}$$

In the two-stage DEA method (hereafter 2-DEA), the efficient frontier and the firm-level efficiency scores are first estimated by DEA or other non-parametric method using a representation of the firm's technology *without* environmental variables, as in (4).

Let $\hat{E}_i$ denote the first-stage estimate of firm's (x_i,y_i) efficiency level. In the second stage, the estimated DEA efficiency scores are regressed on contextual variables. The two-stage regression can be written in general terms as:

$$\hat{E}_i=\tau z_i+\varepsilon_i\geq 1, \tag{55}$$

where τ is a vector of parameters, and ε_i is a random variable. The inequality in (55) yields a truncated (linear) regression model.[53]

From the equations above, it is straightforward to notice that the main conceptual difference between one and two-stage methods is that the one-stage methods incorporate the z-variables as frontier determinants, whereas the two-stage methods incorporates the z-variables as determinants of firms' inefficiency, which in turn is measured with respect to an uncorrected production (cost) frontier. This difference implies that the sign of the contextual variables is assumed to be known beforehand in one-stage DEA methods, whereas the sign of these variables is estimated in two-stage methods.[54] Thus, from a conceptual point of view, 2–DEA methods are more appropriate in applications where the environment is multifaceted and consists of a large and varied number of factors with complex interactions, so that it is difficult to formulate hypotheses with respect to the effect of weather conditions on firms' performance.

The choice of a proper method to control for environmental conditions has attracted merited attention in the DEA literature. The seminal paper of Banker and Morey (1986) modified the measure of inefficiency obtained by removing the effect of *contextual* variables on the measured inefficiency level within the DEA model. Ruggiero (1996) and other authors have highlighted that the one-stage model introduced by Banker and Morey (1986) might lead to bias because the modified DEA model does not properly reflect the importance that environmental variables have on production, and may overestimate the level of technical inefficiency. To solve this problem,

[52] Although the two-stage method is the most popular method in DEA for identifying inefficiency determinants, three-stage models have also been developed (see, for instance, Fried et al. 2002; and Muñiz, 2002).

[53] Interesting enough, this specification of the way efficiency scores depend on z-variables corresponds to the popular KGMHLBC model in the SFA approach (see next subsection).

[54] In a 1–DEA method, one must not only decide a priori what the role of z is, but also free disposability is assumed when FDH and DEA techniques are used.

other models using several stages have been developed in the literature. Ray (1988) was the first to propose a second stage where standard DEA efficiency scores were regressed on a set of contextual variables. The two-stage DEA method has proved useful in a large number of applications where a variety of regression techniques have been used, including the traditional ordinary least squares, censored (tobit and probit) regression, or log-normal transformations.

The 2–DEA method was widespread until Simar and Wilson (2007) demonstrated that this procedure is inconsistent because it lacks a coherent data generating process and the first-stage DEA efficiency estimates are serially correlated. Most notably, they state that the conventional methods of statistical inference are invalid in the second stage regression. The problems arise from the fact that (55) is the assumed model, whereas the true model is

$$E_i = \tau z_i + \varepsilon_i \geq 1 . \tag{56}$$

The dependent variable in (56) is not observed and must be replaced by an estimate $\hat{E}_i$. Simar and Wilson (2007) show that unfortunately $\hat{E}_i$ is a biased estimator of E_i because, by construction, z_i is correlated with the error term ε_i.

To address this issue, these authors propose the use of a bootstrap method to correct for the small sample bias and serial correlation of the DEA efficiency estimates. Further, they advocate the use of the truncated regression model that takes into account explicitly the bounded domain of the DEA efficiency estimates.[55] Since this remarkable paper, the statistical foundations of the 2–DEA method have been subject to intensive debate. For instance, Banker and Natarajan (2008) show that the second-stage OLS estimator of the contextual variables is statistically consistent under certain assumptions and regularity conditions. They also report Monte Carlo simulations that indicate that the 2–DEA method performs reasonably well when the contextual variables are uncorrelated with the inputs. However, as the correlation between the inputs and contextual variables increases, the performance of the 2–DEA estimator deteriorates.

Subsequent discussion has focused on the assumptions. Banker and Natarajan (2008) argue that their statistical model requires less restrictive assumptions than the model of Simar and Wilson (2007). Later on, Simar and Wilson (2010) outlined a list of seven assumptions in the Banker and Natarajan paper that they find restrictive. Johnson and Kuosmanen (2012) further elaborate the assumptions and the statistical properties of the two-stage estimators under more general assumptions than those imposed by Banker and Natarajan (2008). In particular, they show that consistency of the 2–DEA estimator does not require that the contextual variables are uncorrelated with inputs. Unlike SFA regression methods, the DEA efficiency estimator is not subject to the omitted variable bias in the first stage if the effect of the contextual variables has a finite maximum and the sample size is sufficiently large. However, the small sample bias of the DEA estimator will carry over to the second stage regression.

Johnson and Kuosmanen (2012) also develop a new one-stage semi-nonparametric DEA-style estimator that facilitates joint estimation of the frontier and the effects of contextual variables. Their 1–DEA method suggests that unbiased and efficient estimation of the effects of contextual variables *requires* joint estimation of the frontier and the coefficients of the contextual variables. Interestingly enough, to develop such an estimator, they apply insights from their earlier study

[55] Daraio and Simar (2005) proposed an alternative approach by defining a conditional efficiency measure. This approach does not require a separability condition as required by the two-stage approach.

(Kuosmanen and Johnson, 2010), where they showed that standard DEA has a regression interpretation. The main advantage of their 1–DEA method is that it takes into account the correlation of inputs and contextual variables in the simultaneous estimation of the frontier and the effects of contextual variables. In Johnson and Kuosmanen (2012), they introduce the contextual variables to the so-called StoNED model, where the *z*–variables are incorporated additively to the parametric part of the model, which is estimated jointly with the nonparametric frontier. The new StoNED method is similar to the 1–DEA in that it jointly estimates the frontier and the contextual variables using convex nonparametric least squares regression. Both models mainly differ in the assumption made with respect to the truncated noise term.

In the recently developed semi-parametric literature, it is worthwhile mentioning another two models that also allow controlling for environmental variables. The first one is the Semiparametric Smooth Coefficient Model (SPSCM) introduced by Li et al. (2002) where the regression coefficients are unknown functions, which depend on a set of contextual variables. Sun and Kumbhakar (2013) extend this model by allowing the environmental variables to also enter through the inefficiency. The second model is the latent class model (LCM), where *z*–variables enter in non-linear form for the probabilities of belonging to the classes (see e.g., Orea and Kumbhakar, 2004).

5.2. The inclusion of contextual variables in SFA

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*. Parmeter and Kumbhakar (2014) provide a detailed discussion on the consequences of ignoring exogenous determinants of the specific level of inefficiency for a given firm (or heteroscedasticity in the random inefficiency term). It has long been recognized that such a two-step procedure will give biased results. For instance, Wang and Schmidt (2002) find serious bias at all stages of this procedure using Monte Carlo evidence. As the size of the bias is very substantial, they argue strongly against two-step SFA procedures. The solution to this bias problem is a one-step procedure based on the correctly specified model for the distribution of *y* given *x* and *z*.

5.2.1. Frontier determinant vs. determinant of firms' inefficiency

The first methodological choice is again whether we should incorporate the *z*-variables as either frontier determinants (as in the one-stage DEA method), determinants of firms' inefficiency (as in the two-stage DEA method), or as determinants of both the frontier and the inefficiency term. While the above dilemma may not be very relevant in practice as the sign of the contextual variables is not necessarily assumed to be known beforehand using a SFA approach, the key question that should be responded in order to include the *z*–variables as frontier determinants is whether a fully (100%) 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. To respond properly to this question most likely requires having a good knowledge of the industry that is being examined, or recurring to technical (e.g., engineering) support.

Whether *z*–variables should be included in the frontier function or the inefficiency term may *not* be a semantic issue from a conceptual point of view, and might have very different implications for policy makers, regulators and managers. For instance, the traditional time trend can be viewed as a non-controllable *z*–variable. If this variable is added to the production or cost frontier, it captures technical change. A poor rate of technical progress might suggest implementing policy measures encouraging R&D activities. In contrast, if the same variable is included as a determinant of firms' inefficiency, it captures changes in firms' inefficiency over time. In this case, deterioration in firms' performance might suggest implementing policy measures aiming to improve (update) managerial skills.

The above distinction may also be important in regulated industries where regulators purge firms' cost data in order to control for differences in environmental conditions. In these settings, it might be contentious whether we should purge the data regardless of whether the environmental variables are part of the technology (i.e., they are frontier drivers independent from firms' performance), or they have an indirect effect through inefficiency (indicating that, for instance, it is more difficult to manage firms operating in regions with adverse weather conditions). That is, conditional on the technology (that might already include *z*–variables), can firms use unfavourable weather conditions as an excuse to avoid being penalized due to their bad performance? As the environment is not controlled by the firm, one might argue that firms should not be blamed for environment-induced inefficiency. This interpretation implies that regulators should purge firms' cost data when environmental conditions have both direct and indirect effects on firms' cost. We should point out, however, that purging the data completely is likely to be a fairer policy in the short-run, i.e., conditional on current firms' managerial skills. However, if the estimated indirect effect is significant, one could conclude that not compensating all weather effects could help to encourage these firms to hire better qualified executives and staff, perhaps not immediately, but at least in the long-run. Thus, regulators might be aware of this trade-off between short and long-run objectives when they design their incentive schemes.

5.2.2. How to incorporate *z*–variables as inefficiency determinants

A second methodological choice appears when the *z*–variables are treated as inefficiency determinants and hence heteroscedastic SFA models are to be estimated. Summaries of this literature can be found in Kumbhakar and Lovell (2000) and Parmeter and Kumbhakar (2014). A comprehensive heteroscedastic SFA model that nests another heteroscedastic SFA model is the general exponential model (GEM) introduced by Alvarez et al. (2006), that can be written as:

$$y_i = x_i'\beta + v_i - u_i, \tag{57}$$

where $x_i'\beta$ is the log of the frontier production (distance) function (e.g., Translog), $u_i \sim N^+\left(\mu_i, \sigma_{ui}^2\right)$, $\mu_i = \exp\left(\delta_0 + z_i'\delta\right)$, $\sigma_{ui} = \exp\left(\gamma_0 + z_i'\gamma\right)$, and δ_0, δ, γ_0 and γ are parameters to be estimated, and z_i is a vector of efficiency determinants.

The environmental variables enter into the GEM model both through the pre-truncation mean and variance of the inefficiency term, and hence the model allows for non-monotonic effects of the *z*-variables on firms' inefficiency (see Wang and Schmidt, 2002). According to this model, most heteroscedastic SFA models can be divided into three groups: i) the KGMHLBC model, for Kumbhakar *et al.* (1991), Huang and Liu (1994), and Battese and Coelli (1995); ii) the RSCFG

model, for Reifschneider and Stevenson (1991), Caudill and Ford (1993), Caudill et al. (1995); and iii) the general models introduced by Wang (2002), Alvarez et al. (2006) or Lai and Huang (2010).

The above-mentioned models differ on how the contextual variables are introduced, i.e., through the mean of the normal distributed random variable that is going to be truncated over zero; through pre-truncated variance, or simultaneously through the pre-truncated mean and variance. Indeed, the contextual variables are introduced in the general models through both the mean and variance of the normal distributed random variable. In the KGMHLBC models, it is assumed that the variance of the pre-truncated normal variable is homoscedastic (i.e., γ=0) and, thus, the contextual variables are introduced here through the pre-truncated mean.[56] In contrast, in the RSCFG model, it is assumed that the mean of the pre-truncated normal variable is homoscedastic (i.e., δ=0) and, hence, the environmental variables are treated as determinants of the variance of the pre-truncated normal variable.

5.2.3. Models possessing the scaling property

The original RSCFG-type models proposed in the literature assumed in addition that μ=0. As a consequence of this assumption, the so-called *scaling property* is satisfied in this model in the sense that the inefficiency term can be written as a deterministic (scaling) function of a set of efficiency covariates times a one-sided random variable that does not depend on any efficiency determinant. That is:[57]

$$u_i = h_i(z_i, \gamma)u_i^*, \quad u_i^* \sim N^+\left(\mu, \sigma_u^2\right). \tag{58}$$

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.

---

[56] While the KGMHLBC model parameterizes the pre-truncation mean of the distribution as a linear function of the z-variables (see also Wang, 2002; and Lai and Huang, 2010), the parameterization in the GEM model is done by an exponential function of z. Although one parameterization is just a simple monotonic transformation of the other, the econometric implications of the two specifications are quite different. It is because the pre-truncation mean of the distribution is bounded to be positive using the exponential function, which can be quite restrictive for empirical purposes. However, these authors failed to obtain results using many different data sets when trying to estimate a model with a linear specification of the pre-truncation mean. Lack of convergence is a frequent outcome when estimating SFA models due to the likelihood function being highly non-linear. Without the aim of opening a methodological discussion here, we feel that the lack of convergence in these models could be caused by the fact that the pre-truncated mean is likely negative for some observations. In these cases, the distribution of the inefficiency term tends to be more symmetric, and this does not help to identify the one-sided error term.

[57] In a panel data setting, the inefficiency term can be decomposed as $u_i^t = h_i^t(z_i^t, \gamma)u_i^*$ (see Battese and Coelli, 1992) or as $u_i^t = h_i^t(z_i^t, \gamma)u_{it}^*$ (see Álvarez et al. 2006).

The KGMHLBC model and Wang (2002) do not have this scaling property. 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 (both with exponential functions, say) and with the same parameters in each parameterization. In this sense, another model that also satisfies the scaling property is the so-called *scaled Stevenson* model introduced by Alvarez *et al.* (2006). In this model, both the mean and the variance of the pre-truncated normal depend on the contextual variables but the coefficients of the contextual variables in the pre-truncation mean and variance in (57) are the same, i.e., $\delta = \gamma$. Moreover, Parmeter and Kumbhakar (2014) also mention that any distributional assumption involving a single parameter family (such as half-normal or exponential) will automatically possess the scaling property.

Although it is an empirical question whether or not the scaling property should be imposed, and not all commonly used models fulfil this property, it has some features that make it attractive to some authors (see Wang and Schmidt, 2002). As noted by Simar et al. (1994) and Wang and Schmidt (2002), perhaps the most fundamental statistical benefit of the scaling property is that the stochastic frontier and the deterministic component of inefficiency can be recovered without requiring a specific distributional assumption on u_i, such as half-normal or truncated half-normal. Indeed, if we take into account our specification of firms' inefficiency in (58) and define $u^* = E(u_i^*)$, then taking expectations in (57) yields:

$$y_i = x_i'\beta - h_i(z_i,\gamma)\,u^* + \varepsilon_i^*, \tag{59}$$

where $\varepsilon_i^* = v_i - h_i(z_i,\gamma)\left[u_i^* - u^*\right]$. Equation (59) can be estimated by non-linear least squares (NLLS). The need for NLLS stems from the fact that the scaling function must be positive. Given that ε_{it}^* is heteroscedastic, generalized NLLS would be needed to obtain estimates that are efficient (for the class of estimates that do not impose distributional assumptions). In addition, robust standard errors should be constructed to ensure valid inferences. As Parmeter and Kumbhakar (2014) point out, the model in (59) can be viewed as a (restricted) version of the partly linear model of Robinson (1988). These authors show that, if z_{it} and x_{it} do not include common elements, the conditional mean $E(u_i|z_i)$ can be estimated in a nonparametric fashion without requiring distributional assumptions for u_i. The frontier parameters can be estimated by conditioning equation $E\left[y_i \mid z_i\right]$ from (59). This yields the following model:

$$y_i - E\left[y_i \mid z_i\right] = \left(x_i - E\left[x_i \mid z_i\right]\right)'\beta + \varepsilon_i^*. \tag{60}$$

If $E\left[y_i \mid z_i\right]$ and $E\left[x_i \mid z_i\right]$ were known, the frontier parameters could be estimated by OLS. In practice, both conditional means are replaced in the model of Robinson (1988) with nonparametric estimates. $E(u_i) = h_i(z_i,\gamma)u^*$ is estimated later on via local least squares. It should be noted that whereas the scaling property was required for parametric estimation of $E(u_i|z_i)$, no such restriction is needed in the partial nonparametric model introduced by Parmeter et al. (2016). The scaling property is only imposed (ex post) through the interpretation of $E(u_i)$.

Several authors have found the scaling property useful for other purposes. For instance, Wang and Ho (2010) also used a multiplicative decomposition of the inefficiency term in order to control for unobserved heterogeneity in a panel data setting. Let us consider that the I firms are observed in t=0,…,T periods, add a fixed effect to (57) and assume that while the scaling function

in (58) varies over time, the random inefficiency term is time invariant, as in Battese and Coelli (1992). In this case, the model to be estimated can be written as:

$$y_i^t = \alpha_i + x_i^t{}'\beta + v_i^t - h_i^t(z_i^t,\gamma)\,u_i^* \ . \tag{61}$$

If one treats α_i as a random variable that is correlated with x_i^t but does not capture inefficiency, then the above model becomes what has been termed the "true fixed effects" panel stochastic frontier model introduced by Greene (2005). The model is labelled as the "true random effects" model when α_i is treated as uncorrelated with x_i^t. Estimation of the model in (61) is not straightforward because the incidental parameters problem arises when the number of parameters to be estimated increases with the number of firms in the data. Wang and Ho (2010) solve this problem by proposing a class of stochastic frontier models in which the within and first-difference transformation of the model can be carried out while also providing a closed form likelihood function. For instance, a first-difference transformation of (61) yields the following equation to be estimated:

$$\Delta y_i^t = \Delta x_i^t{}'\beta + \Delta v_i^t - \Delta u_i^t = \Delta x_i^t{}'\beta + \Delta v_i^t - \Delta h_i^t(z_i^t,\gamma)\,u_i^* , \tag{62}$$

where Δ stands for the first-difference transformation of the variables. The main advantage of this specification of the model is that because the fixed-effects are removed from the model, the incidental parameters problem is avoided entirely. Note that, although the distribution of Δv_i^t has a closed form, the distribution of Δu_i^t is *generally* not known *if* we assume that u_i^t is independently distributed across firms (see, for instance, Wang, 2003). However, if the inefficiency term possesses the scaling property and the basic inefficiency is time invariant, the distribution of u_i^* is not affected by the first-difference transformation. As Wang and Ho (2010) point out, this key aspect of the model leads to a tractable likelihood function.

As we will see later on Section 6, Griffiths and Hajargasht (2016) also take advantage of the scaling property to address endogeneity issues. To model correlation between the inefficiency error u_i^t and some or all of the inputs they assume that there is a transformation of u_i^t, call it $H(u_i^t)$, that yields the following specification of the transformed inefficiency term:

$$H(u_i^t) = x_i^t{}'\gamma + e_i^t , \tag{63}$$

with e_i^t being a normal distributed random term. Note that equation (63) is similar to (58)—appending the time superscript—if the transformation is the logarithmic one; i.e., $H(u_i^t)=\ln(u_i^t)$. Thus this endogeneity model can be viewed as a stochastic frontier model with scaling properties where the set of z-variables includes (partly or completely) input variables. In this sense, it is worth mentioning that many practitioners (still) think that the set of inefficiency determinants cannot include any input variable in order for the model to be well specified. This contribution shows that the key issue is not whether z and x matrices overlap, but whether the input variables are correlated with the basic inefficiency term (here e_i^t).

Finally, Orea et al. (2015) brings attention to the fact that the energy demand frontier model introduced by Filippini and Hunt (2012) is closely connected to the measurement of the so-called rebound effect associated with improvements in energy efficiency.[58] Moreover, they show

[58] The rebound effect is a phenomenon associated with energy consumption. This concept has to do with the idea that an increase in the level of efficiency in the use of energy decreases the marginal cost of supplying a certain energy service and hence may lead to an increase in the consumption of that service. This consumer reaction might therefore

that a traditional demand frontier model implicitly imposes a zero rebound effect. They next take advantage of the scaling property to relax this restrictive assumption.

As the rebound effect tends to attenuate, exacerbate, or even reverse the effect of improvements in energy efficiency on energy consumption,[59] this effect can be introduced in an energy demand application of the SFA approach as a *correction factor* that interacts with the energy inefficiency term (u_i) that is appended to the stochastic energy demand frontier. That is:

$$\ln q_i = f_i(\cdot) + v_i + (1 - R(z_i, \gamma))\, u_i = f_i(\cdot) + v_i + h(z_i, \gamma)\, u_i , \quad (64)$$

where $R(\cdot)$ is a function measuring the rebound effect that depends on a set of energy services determinants likely associated with this (also unobserved) determinant of the energy demand.. Therefore, their empirical strategy relies on a convenient reinterpretation of the stochastic frontier model that satisfies the traditional scaling property in production economics. Like the SFA models in production economics, only u_i is treated here as (energy) inefficiency. The so-called scaling function is used in this case as a measure of the rebound effect.[60]

6. Endogenous issues in frontier models

6.1. Alternative methods to deal with endogenous regressors

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 the regressors (e.g., inputs and/or outputs levels). Early papers such as Mundlak (1961) and McElroy (1987), and Kumbhakar and Tsionas (2011) pointed out that the regressors are likely endogenous when some inputs or outputs are chosen by the firms to maximize or minimize some objective function *and* the random shocks and firms' inefficiencies

partially or totally offset the predicted reduction in energy consumption attributed to energy efficiency improvements using engineering models.

[59] It is not easy to find a similar phenomenon in efficiency and productivity studies where SFA models have traditionally been applied. In that literature any improvement in firms' efficiency is assumed to have a proportional effect on firms' performance (outputs, cost, etc.). Just to conjecture an example, a sort of rebound effect might appear in public firms where employees' salary is not linked to their productivity. In this case, an employee who works efficiently could become "lazy" after a salary improvement since his earnings do not depend on his effort. Another example of rebound effect may also occur when labour productivity increases but this does not lead to one-for-one reductions in employment.

[60] Orea et al. (2016) propose exploring two simple rebound-effect functions: whereas $R(z_i, \gamma) = (e^{z_i'\gamma} - 1) / e^{z_i'\gamma}$ can depict any value lower than full rebound and even allows to obtain *super-conservation* (SC) outcomes (i.e., R<1), the rebound-effect function $R(z_i, \gamma) = e^{z_i'\gamma} / (1 + e^{z_i'\gamma})$ precludes this somewhat counter-intuitive outcome as it only allows for *partial* (PA) rebound-effects (i.e., 0<R<1). The PA rebound effect model implies a scaling function that is similar to the logistic scaling function in Kumbhakar (1991). With the implied scaling function in the SC rebound effect model, the term can be referred to the exponential scaling function in Battese and Coelli (1992).

are observed by the producer (but unobserved by the researcher).[61] Later on, using distance functions, in a cost minimization setting Tsionas et al. (2015) show that the input ratios in an input distance function are endogenous if allocative errors are correlated with technical inefficiency and/or random productivity shocks. A similar conclusion can be obtained for the output ratios in an output distance function if firms maximize revenues. In accordance with the previous statements, both inputs and outputs are endogenous if firms maximize profits. 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 by the market.

Although endogeneity issues were first discussed in the regression framework, it has also been addressed in the programming approach. Therefore, in this section we present a series of models addressing endogeneity in the non-parametric DEA and parametric SFA frameworks, and correcting for the likely biases in the regressors and decision variables−weights−determining efficiency levels.

6.1.1. Endogeneity in DEA models.

It should be mentioned that statistical issues linked to endogeneity do not arise explicitly when the frontier technology is estimated through the deterministic mathematical programming approach. However, ignoring the random nature of the data generating process does not preclude the existence of endogeneity problems when a DEA or other non-parametric technique is used. Orme and Smith (1996), Peyrache and Coelli (2009) and Santín and Sicilia (2017) discuss endogeneity problems in the calculation of efficiency in DEA models when the regressors (inputs or outputs) are correlated with technical inefficiency.

Initially, Wilson (2003) surveys a number of tests that can be used to determine the independence and uncorrelated hypotheses in the context of efficiency measurement, including their merits and drawbacks. The alternative possibilities are based on analytical and asymptotic results and refer to: i) nonparametric tests for independence determining if the marginal probability densities of the efficiency and the regressors—$f(\zeta)$ and $f(x)$—verify the following relation with their joint density function: $f(\zeta, x) = f(\zeta) f(x)$; ii) smoothing methods considering kernel estimates (Ahmad and Li, 1997; Zheng, 1997); iii) correlation integrals—Johnson and McClelland (1998); and iv) ranks such as the usual Spearman's ρ or Kendall's τ tests—see Kallenberg and Ledwina (1999) generalizing them through the use of cupolas. Afterwards, this author performs Monte Carlo simulations to establish that these tests have poor size properties and low power in moderate sample sizes. Peyrache and Coelli (2009) build upon the previous findings and propose a semi-parametric Hausmann-type asymptotic test for linear independence (uncorrelation), and also resorting to Monte Carlo experimentation, show that it has good size and power properties in finite samples.

Additionally, Cordero et al. (2015) show that the effect of endogeneity on DEA efficiency estimates and the extent to which they deviate from their true (designed) values is positively related to the level of correlation between the regressors and the efficiency. They show that with low and

[61] On the other hand, Zellner et al. (1966) took the opposite view and assumed that the inefficiency term is unknown to both producers and analysts. Under this scenario, there is no endogeneity problem in estimating production function by OLS.

moderate levels of correlation, the standard DEA model performs well, but that for high levels—either negative or positive, it is mandatory to use instrumental techniques that correct the bias. Therefore, in empirical applications, when the existence of endogeneity is suspected, it becomes mandatory to identify it by applying non-parametric tests, and then correct it with suitable techniques.

Based on these findings Santín and Sicilia (2017) devise a semi-parametric strategy similar to the instrumental variables (IV) approach in regression analysis—discussed in the following section, and that results in a DEA specification that accounts for the exogenous part of the endogenous regression and that is uncorrelated with technical efficiency. As in the parametric counterpart, the first step is to choose an instrumental variable z that is significantly correlated with the endogenous regressor x, i.e., $E(x \mid z) \neq 0$ (relevance), but uncorrelated with the true efficiency, i.e., $E(\zeta \mid z) = 0$ (exogeneity). Empirically, the relevance condition can be tested regressing the endogenous regressor x on the exogenous regressors, x^{ex} and the instrument z,

$$x = \alpha + \beta x^{ex} + \chi z + \varsigma, \tag{65}$$

and testing the significance of its associated parameter ($H_0 : \chi = 0$). If H_0 is not rejected, the instrument is relevant. As for the second condition of exogeneity it cannot be tested since it is unobserved. In that case these authors suggest interpreting it as the absence of correlation between the instrument z and the variables characterizing the alternative dimension of the production process, e.g., outputs y in the case of endogenous inputs. In that case, z should not have a partial effect on y (beyond its effect on the endogenous input) and should be uncorrelated with any other omitted variables (when this is the cause of endogeneity). Under these assumptions, the authors implement an instrumental variable process by substituting the estimated exogenous regressor —input $\hat{x}$— from (65) for the endogenous regressor when solving the DEA program associated to the relevant orientation, e.g., output.

Again, relying on Monte Carlo experimentation they generate efficiency values from the Cobb-Douglas and translog specifications with different samples sizes (50, 100, and 300) using a benchmark data generating process with no correlation, and different levels of positive and negative correlation between the true efficiency and the endogenous input regressor: low ($\rho = \pm$ 0.2), medium ($\rho = \pm$ 0.4) and high ($\rho = \pm$ 0.8). Considering alternative indicators to measure the difference between the true and estimated efficiency, such as Spearman's ρ and Mean Absolute Deviation (MAD), they find that both standard and instrumental DEAs yield similar results in the case of low correlation, still with marked differences with respect to the true efficiency values, but that the latter clearly outperforms the former under high correlation $\rho = 0.8$. They also confirm Cordero et al.'s (2015) results with standard DEA being robust for negative correlations. Also, coinciding with previous studies the instrumental DEA performs better as the sample size increases (Orme and Smith, 1996; Krüger, 2012).

In sum, it has been established that under the suspicion of regressor endogeneity, new methods are available to test its existence and, then, correct the bias it creates through suitable instrumentation, thereby improving the reliability of the results that inform decision making processes. Nevertheless, this area of research lags behind the methods devised to correct endogeneity issues in the parametric SFA approach.

6.1.2. Endogeneity in SFA models.

Researchers need to deal with endogeneity issues because the usual procedures for estimating SFA models depend on the assumption that the inputs are exogenous. However, dealing with the endogeneity issue is relatively more complicated in a stochastic frontier analysis framework than in standard regression models due to the special nature of the error term (Karakaplan and Kutlu, 2015). To start with, it should be mentioned that it is not appropriate to insert ''fitted values'' for the endogenous variables and then proceed with standard SFA procedures (Amsler et al, 2016).

Endogeneity is often handled in standard regression models by using instrumental variables estimators (e.g., 2SLS or GMM), or estimating by MLE a system of equations that contains the equation of interest plus the reduced form equations that link the endogenous variables with the set of instruments. While adapting an IV-GMM procedure to the SFA framework model is straightforward, modifying a MLE stochastic frontier model is not because it is not clear how best to model the joint distribution of the composed error in the stochastic frontier and the error in the reduced form equations for the endogenous variables.

Several authors have recently proposed alternative empirical strategies to account for endogenous regressors in SFA settings. Some of them allow only for correlations between the regressors and the noise term (e.g., Tran and Tsionas, 2013), while other authors allow for correlations with the inefficiency term (e.g., Amsler et al, 2016). Models can be estimated using IV-GMM techniques (e.g., Guan et al., 2009), ML procedures (e.g., Kutlu, 2010) or Bayesian estimation methods (e.g., Griffths and Hajargasht, 2016).[62] Moreover, many of them can be estimated in one or two-stages. Therefore, the researcher has several methods at hand to deal with endogeneity issues when estimating a SFA model. 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 and panel data:

$$\begin{aligned} \ln y_i^t &= x_i^t{}'\beta + v_i^t - u_i^t, \\ x_i^t &= z_i^t{}'\delta + \eta_i^t, \end{aligned} \tag{66}$$

where x_i^t is a px1 vector of endogenous variables (excluding $\ln y_i^t$), and z_i^t is a qx1 vector of exogenous or instrumental variables, and the second equation in (66) can be viewed as a reduced form equation that links the endogenous variables with the set of instruments. The endogeneity problem arises if η_{it} in the second equation is correlated with either v_{it} or u_{it} in the first equation.

6.1.2.1. GMM estimation

In order to estimate consistently the frontier model (66), Guan et al. (2009) propose a two-step MM estimation strategy. In the first step, they ignore the structure of the composed error term and suggest estimating the frontier parameters using a GMM estimator as long as valid instruments are found. Suppose that the vector of instruments z_{it} satisfies the following moment condition:

$$E[z_i^t\,(v_i^t - u_i^t)] = E[z_i^t \cdot (\ln y_i^t - f(x_i^t, \beta))] = E[m_i^t(\beta)] = 0. \tag{67}$$

[62] See also Olley and Pakes (1996), Levinsohn and Petrin (2003), and Wooldridge (2009) on the solution to the endogeneity problem in non-frontier applications using panel data without using any behavioural assumption.

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

$$\hat{\beta} = \arg\min\left[\Sigma_i \Sigma_t m_i^t(\beta)\right]' \Omega^{-1} \left[\Sigma_i \Sigma_t m_i^t(\beta)\right] = 0, \tag{68}$$

where Ω is an optimal weight matrix obtained from a consistent preliminary GMM estimator. This optimal weight matrix should take into account the heteroscedasticity of the error term. In the second step, distributional assumptions are invoked to obtain ML estimates of the parameter(s) describing the variance of v_i^t and u_i^t, conditional on the first-stage estimated parameters. The ML estimates of σ_v and σ_u are obtained by maximizing the likelihood function associated to the following error term:

$$e_{it} = \ln y_i^t - x_i^t{}' \hat{\beta}, \tag{69}$$

where e_i^t are the residuals from the first-step estimation. An important issue to consider is that the zero-mean residual e_i^t is not equal to $\varepsilon_i^t = v_i^t - u_i^t$ because u_i^t is a non-negative random term. If the inefficiency term follows a homoscedastic distribution, the GMM intercept is biased, and the bias is equal to the expectation of the original error term in (66). This implies that the estimated value of the error term in equation (69) should be decomposed as follows:

$$e_i^t = v_i^t - u_i^t + E(u_i^t). \tag{70}$$

Note that no new parameters are to be estimated if, for instance, u_i^t follows a half-normal distribution. In this case the expected value of u_i^t depends on the standard deviation of the pre-truncated normal variable. One advantage of this approach is that the stochastic frontier model based on (70) can accommodate both *heteroskedastic* inefficiency and noise terms simply by making the variances of σ_v and σ_u functions of some exogenous variables, an issue that has been discussed in the previous section (see, for instance, Wang, 2002; Álvarez et al., 2006). However, we should be aware that ignoring that in the first-stage of the process the inefficiency term depends on a set of covariates could bias *all* model parameters. Indeed, an important 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. Guan et al. (2009) mention this issue, but do not discuss its implications for the GMM estimation. Indeed, to achieve consistent estimates it is critical to ensure that the chosen instruments do not include determinants of u_i^t.

Kumbhakar et al. (2013) and Malikov et al. (2015) suggest bringing economic behaviour into the analysis to solve endogeneity problems. They address the endogeneity of output and input variables assuming that firms respectively maximize profitability (or return to the outlay, RO) or minimize cost. Instead of introducing instruments for these endogenous variables in an *ad hoc* fashion (e.g., temporal lags of inputs and outputs), they address the endogeneity issue by defining a system in which they bring additional equations for the endogenous variables from the first-order conditions of RO (cost) maximization (minimization). Their first-order conditions can be viewed as reduced form equations where the price variables are assumed to be exogenous. 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). Their model can also accommodate both technical and (input) allocative inefficiencies among firms. 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.

Kumbhakar (2011) also relies on economic behaviour, but in this contribution he solves the endogeneity of both outputs and inputs first by deriving a particular form of the estimating equation in which the regressors are ratios of inputs and outputs. He shows that these ratios are uncorrelated with the error term in the estimating model if producers maximize return to the outlay. Thus his transformed equation can be estimated consistently by ML methods using standard stochastic frontier software.

Tran and Tsionas (2013) propose later on a GMM variation of the ML-based model introduced by Kutlu (2010). They interpret Kutlu's (2010) first order conditions as moment conditions and propose a simple GMM estimator that is consistent and asymptotically efficient. It provides consistent and correct standard errors of the estimated parameters, and it is fairly simple to compute given the current existing computing power and readily automated GMM estimation programs. In this vein, Amsler et al. (2016) suggest using the GMM approach of Hansen et al. (2010) and replace one of the moment conditions defined under exogeneity in a standard SFA model with an equivalent moment condition that uses the set of instruments z_{it}. This approach requires however independence of z_{it} and ε_{it}.

6.1.2.2. ML estimation

Kutlu (2010), Tran and Tsionas (2013) and Amsler et al. (2016) make efforts to address the endogeneity problem in a fully maximum likelihood 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 (66). 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 inefficiency term.

Kutlu (2010) proposes a model that aims to solve the endogeneity problem due to the correlation between the regressors and the two-sided error term.[63] He assumes that the error terms ε_{it} and v_{it} in (66) satisfy the following:

$$\begin{pmatrix} \Omega^{-1/2}\eta_i^t \\ v_i^t \end{pmatrix} = N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \mathbf{I}_p & \rho\sigma_v \\ \rho'\sigma_v & \sigma_v^2 \end{bmatrix}\right) \tag{71}$$

where Ω_η is the variance-covariance matrix of η_{it} and ρ is a correlation vector between v_i^t and η_i^t. Based on (71), the equations in (66) can be written as (see Tran and Tsionas, 2013; p. 234):

$$\ln y_i^t = x_i^t\beta + \tau(x_i^t - z_i^t\delta) + \omega_i^t - u_i^t, \tag{72}$$

where $\omega_i^t=(1-\rho'\rho)v_i^t$ and $\tau = \sigma_v\rho'\Omega^{-1/2}$, which can be viewed as a correction term for bias. Note that $\omega_{it}+u_{it}$ is conditionally independent from the regressors given x_i^t and z_i^t. Hence, conditional on x_{it} and z_{it}, the distribution of the composed error term in (72) is exactly the same as their traditional counterparts from the stochastic frontier literature. They then show that for the sample observations (y_i^t, x_i^t, z_i^t), the joint log-likelihood function of y_i^t and x_i^t is given by

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

$$\ln L(\theta) = \ln L_{y|x}(\theta) + \ln L_x(\theta), \tag{73}$$

where

$$\ln L_{y|x}(\theta) = -\frac{IT}{2}\ln\left(\sigma_\omega^2 + \sigma_u^2\right) + \frac{1}{\left(\sigma_\omega^2 + \sigma_u^2\right)^{1/2}}\sum_{i=1}^{I}\sum_{t=1}^{T}\ln\Phi\left(-\left(\ln y_i^t - x_i^t\beta - \tau(x_i^t - z_i^t\delta)\right)\sigma_u/\sigma_\omega\right)$$
$$-\frac{1}{2\left(\sigma_\omega^2 + \sigma_u^2\right)}\sum_{i=1}^{I}\sum_{t=1}^{T}\left(\ln y_i^t - x_i^t\beta - \tau(x_i^t - z_i^t\delta)\right)^2, \tag{74}$$

and

$$\ln L_{y|x}(\theta) = -\frac{IT}{2}\ln\left(|\Omega|\right) - \frac{1}{2}\sum_{i=1}^{I}\sum_{t=1}^{T}\eta_i^t\Omega^{-1}\eta_i^t\,. \tag{75}$$

The first part of the log-likelihood function (73) is almost the same as that of a traditional stochastic frontier model where we have adjusted the residual by the $\tau(x_i^t - z_i^t\delta)$ factor. The second part is just the likelihood function of a multivariate normal variable.

The likelihood function (73) can be maximized to obtain consistent estimates of all parameters of the model. However, if computationally, difficulties appear, one can use a two-step maximum likelihood estimation method. In the first stage, $\ln(\theta)$ is maximized with respect to the relevant parameters. In the second stage, conditional on the parameters estimated in the first stage, $\ln Ly|(\theta)$ is maximized. 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. Alternatively, an analytical approach is possible as pointed out by Amsler et al. (2016; 284).

6.1.2.3. Copula approach

The above mentioned ML model does not address the potential correlation with the inefficiency term, and neither does it assure consistency of parameter estimates when η_i^t is correlated with both v_i^t and u_i^t. Amsler et al. (2016) propose using a copula in order to specify the joint distribution of these three random variables.[64] Their approach in turn allows correlation of v_i^t and u_i^t without environmental variables. They selected a multivariate normal (or ‘‘Gaussian’’) copula. This choice implies that the joint distribution of v_i^t and η_i^t is multivariate normal, which is what is often assumed in the literature. However, this copula does not permit to analytically integrate u_i^t out from the joint density for v_i^t, and η_i^t. For this reason, the parameter estimates should be obtained by maximum simulated likelihood estimation, where the joint density is approximated by taking many draws from the distribution of u and averaging.[65] One obvious difficulty with the approach of this section is the need to specify a copula. Still, the assumption of a Gaussian copula, is at least more general than the assumption that u_i^t is independent from v_i^t and η_i^t because the

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

[65] Applications of simulations to evaluate a likelihood can be found in Greene (2005; 24), Amsler et al. (2016) and Parmeter and Kumbhakar (2014; sections 6 and 7).

Gaussian copula contains the independence copula as a special case (correlations equal to zero). 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 a reduced form equation that links the endogenous variables with the set of instruments. Thus, their approach is useful when it is not possible to find appropriate instruments. They develop a flexible joint distribution of the endogenous regressor and the composed error that can accommodate any degree of dependency between them. This joint distribution is then used to derive the likelihood function. 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).

6.1.2.4. Mundlak transformation approach

Griffiths and Hajargasht (2016) propose a different approach to address endogeneity issues in SFA models. They first consider a panel stochastic frontier model in which correlations between the effects and the regressors are based on a generalisation of the correlated random effects model proposed by Mundlak (1978) and extended by Chamberlain (1984). They show that by transforming the inefficiency term into a normally distributed random term and modelling endogeneity through the mean or covariance of the normal errors, a range of stochastic frontier models with endogeneity can be handled. Their models are estimated by using both maximum simulated likelihood and Bayesian methods. To model correlation between the inefficiency error u_{it} and some or all of the inputs, they assume that:

$$H(u_i^t) = x_i^{t\prime}\gamma + \zeta_i^t . \tag{76}$$

They use a logarithmic transformation function, that is, $H(u_i^t)=\ln(u_i^t)$, which implies that u_{it} has a lognormal distribution. Note that, if we replace the time-varying vector x_i^t with time-invariant firm averages $\overline{x}_i = T^{-1}\sum_{t=1}^{T} x_i^t$, we obtain an extension of the model considered by Mundlak (1978) for a conventional random effects panel data model with correlated effects. Interestingly enough, when $H(u_i^t)=\ln(u_i^t)$, equation (76) can be written as $u_i^t = e^{x_i^{t\prime}\gamma}\, u_{it}^*$, implying that the inefficiency term has the scaling property. This represents a new utility for this interesting property that can be added to that collected by Parmeter and Kumbhakar (2014).

On the other hand, this approach seems to suggest that a simple empirical strategy to deal with regressors that are correlated with the inefficiency term is to use a heteroscedastic SFA model where the scaling function depends on (a set of) the frontier explanatory variables. However, in this case, we cannot interpret these variables as determinants of firms' inefficiency because they are merely capturing the correlation between regressors and the genuine inefficiency term u_{it}^*. Only this term can be interpreted purely as inefficiency.

6.2. The choice of orientation: endogenous and optimal directions

In the standard distance function approach, researchers often choose between input and output oriented measures of firms' inefficiency and estimate the distance function of their choice. Moreover, as discussed in Section 2.4, when we chose estimating a standard dual representation

of firm's technology, we are choosing implicitly a particular orientation to measure firms' inefficiency.

Also, as anticipated in Section 3.2, the primal representations based on the distance functions are functions of the same vector of inputs and outputs. Both the mathematical programming and econometric methods need further qualifications to identify and calculate/estimate the relevant distance functions. In standard DEA, there are models such as the additive formulation that are non-oriented, but traditional oriented models require the specification of different objective functions, either in the envelopment or the multiplier formulations. In SFA one needs to select a set of inputs and/or outputs to impose a particular homogeneity or translation property. This decision implies using a different empirical strategy or orientation to measure firms' inefficiency. As Greene (2008; p.153) notes, the question is which form is appropriate for a given setting? Also, the emergence of new characterizations of the production technology through the directional and generalized distance functions, both in the DEA and SFA approaches, opens a new range of possibilities related to economic behaviour given their flexibility; i.e., through their duality with the profit and profitability functions, or technological criteria; e.g., choosing the direction that minimizes the distance to the frontier. In DEA this flexibility is related to the choice of the directional vector $(-g_x, g_x)$ and parameter α when solving the mathematical programs. In SFA, flexibility is related to the ability to impose the suitable homogeneity and translation conditions, and its effect on econometric issues such as the estimation methods, the existence of endogeneity, etc.

These developments show that the traditional binary choice between an input or output orientation is not the only option, unless it is grounded on the firm's economic objective. Indeed, most researchers back their decision on a previous discussion of firms' economic objectives. Also, the choice of duality framework (summarized in Section 2.4) to perform an overall economic efficiency analyses depends on the specific characteristics of the study. For instance, as the input distance function suggests when referring to the degree by which the current input level exceeds the input requirement for production of a particular amount of outputs, it is natural to associate it to (lack of) cost minimization, resulting in (15). In this case, it is assumed that inputs are the choice variables and/or the firm can reduce them at least in the short run without reducing output. Likewise, as the output distance function suggests when referring to the degree by which output falls short of what can be produced with a given input vector, it is natural to associate this output-oriented function to revenue maximization, (16). In this case, it is assumed that outputs are the choice and adjustable variables. Thus, while the input orientation is intuitive when output is out of control for the firm (e.g., when demand is determined or fixed), the output-orientation is intuitive when the inputs are exogenously determined.

Regarding dual representations of firms' technology, the cost approach is preferred if the output side of the firms is exogenous and non-discretionary, and the opposite is applicable for the revenue side. The choice between profit and profitability (return-to-dollar) is less clear, as both choices are available when both inputs and outputs can be freely adjusted at the discretion of managers. In the short term managers are normally concerned with attaining maximum profit, but it can be argued that the long term viability of a firm critically depends on its ability to remain competitive in the market, with profitability representing an economically weighted (by prices)

measure of productivity. This is particularly true in markets where the degree of competition is large, firms cannot exert market power, and are expected to make economic profit.[66]

Therefore, the choice of orientation should be determined, at least partially, by the capability of firms to adjust their decisions in order to become fully efficient. However, it should be noted that the distance function concept was originally developed to represent the technology using multiple-input and multiple-output data. Kumbhakar (2012) shows that, while the underlying technology to be modelled is the same, the different orientations only provide different sets of theoretical restrictions to identify the frontier parameters to be estimated. This is also clear in a non-parametric context, where the DEA technology represented by (19) is unique, as for the efficient frontier—e.g., (20) , while technical efficiency can be measured considering alternative orientations. Moreover, in the SFA framework 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 and the potential input (output) adjustment can be obtained from both input or output-oriented distance functions. In a similar manner, Orea et al. (2004) and Parmeter and Kumbhakar (2014; section 4.2) show that both output and input oriented inefficiency scores can be computed from an estimated cost function. Thus, if any measure of firms' inefficiency can be estimated using any primal or dual 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. In the non-parametric DEA framework, Kerstens et al. (2012) and Peyrache and Daraio (2012) study how efficiency results critically depend on the choice of orientation. The latter authors study the sensitivity of the estimated efficiency scores to the direction selection. They propose a set of tools including a measure of efficiency (I), which sheds light on the subjective nature of the peer selection process resulting from the choice of an *a priori* direction. Their efficiency measure allows them to aggregate all the directional measures using the density function of the data in order to generate a process characterized by a weighting scheme. If a direction is associated to a large (low) probability, then their measure assigns a larger (lower) weight. They illustrate their method using Italian agriculture data, and compare their measure to absolute distances (e.g., maximum and minimum values of the DDFs), as well as other indicators (e.g., free disposal hull scores). In the parametric SFA setting, Kumbhakar (2010) shows, using panel data on World Health Organization member countries, that efficiency rankings vary substantially depending on whether output vs. input oriented primal specifications are used to model the technology. Kumbhakar and Tsionas (2006) also estimate input and output oriented stochastic frontier production functions, and find that the estimated efficiency, returns to scale, technical change, etc., differ depending on whether one uses the model with input or output-oriented technical inefficiency. Using a dual approach, Orea et al. (2004) estimate cost frontiers under different specifications which assess how inefficiency enters the data generating process using panel data on Spanish dairy farms. The

[66] In the limiting case of perfect competition, maximum profit and profitability are equivalent in the long run, valued at zero and one, respectively, and economic inefficiency cannot exist. Market dynamics abide by the rules of Darwinism, and it is the ability of firms to be more productive what ensures their survival in an ever-changing and uncertain environment, with technological innovation playing a central role. This justifies the relevance of the productivity analyses presented in Section 8, which allows identifying the nature and rate of technological change and efficiency change, with technical change representing a long-run tendency, and efficiency change short-run cyclical variations around it for those firms capable of surviving.

authors show that the different models yield very different pictures of the technology and the efficiency levels of the sector, illustrating the importance of choosing the most appropriate model before carrying out production and efficiency analyses. Similar comments can be made if directional distance functions are used. For instance, Vardanyan and Noh (2006) and Agee et al. (2012) also show that the parameter estimates depend on the choice of the directional vectors.

Second, 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 that is used to estimate output-oriented technical efficiency cannot be applied to estimate input-oriented production functions. They instead use a simulated ML approach to estimate these functions. Similarly, Orea et al (2004) estimated stochastic cost frontier models with output-oriented measures of firms' inefficiency using a non-linear fixed-effect approach. If, in contrast, inefficiency is modelled as a one-sided random term, Parmeter and Kumbhakar (2014) show that a stochastic cost frontier model with output-oriented inefficiency is difficult to estimate without additional restrictions on the technology. The same happens using distance functions if we measure firms' inefficiency using the set of variables that are not involved in the homogeneity restrictions (see Section 6.2.2.).

In this section we discuss the choice of orientation from a modelling perspective in the DEA and SFA approaches, and summarize the most recent proposals related to the rationale underlying different possibilities, including those endogeneizing the orientation, and driven by the data. This last approach emerges in situations in which there is not an economic or managerial rationale to impose a specific goal.

6.2.1. DEA framework

From a DEA perspective, after the introduction of the directional and generalized distance functions by Chambers et al. (1996, 1998) and Chavas and Cox (1999), respectively, several authors have proposed alternative directions to measure efficiency and studied its properties. The flexibility of these functions emanates from the fact that any directional vector and parameter can be chosen, including those corresponding to the traditional input distance function when $(-g_x, g_y) = (-x, 0)$ and $\alpha=0$, and the output distance function when $(-g_x, g_y) = (0, y)$ and $\alpha=1$, respectively.[67] A drawback of the generalized distance function is that it does not treat inputs and outputs asymmetrically, both between and within vectors, assigning different directions to each input and output variable. Therefore, it cannot provide the discretion that might be needed in empirical studies that require an independent treatment of these variables. On these grounds, and as in the SFA section that follows, we focus our discussion mainly on the directional distance function.

Clearly, when choosing an orientation, several criteria are on the list. The first one mirrors the rationale behind the input and output distance functions, by setting the orientation for each DMU equal to the observed amounts of inputs and outputs, $(-g_x, g_y) = (-x, y)$. Färe and Grosskopf (2000a; p. 98) justify the choice on the grounds that it provides a link and symmetry with the

[67] Also, as their input and output particular cases, the directional and generalized distance functions can be adjusted so as to leave out input and output adjustments (reductions and increases, respectively), when those variables are not under the control of the managers or exogenously fixed, as in the non-discretionary variables models introduced by Banker and Morey (1986)—see Section 5.

traditional distance functions as presented above. This implies solving problem (22) substituting the directional vector by the observed amounts, which is the most common approach in empirical applications relying on the DDF. Alternatively, rather than using individual directions for each firm, it is possible to adopt a so-called 'egalitarian' approach assigning the same direction to all firms. An example of such common direction is to take the average input and output mixes: $(-g_x, g_y) = (-\bar{x}, \bar{y})$, or the unit vector $(-g_x, g_y) = (-1, 1)$. Both have the advantage that the direction is neutral. However, the interpretation of the distance function β as efficiency measure is different. When the directional vector is measured in the units of measurement of inputs and outputs; e.g., as with $(-x, y)$ or $(-\bar{x}, \bar{y})$, the efficiency measure corresponds to the fraction of the observed input and outputs amounts that are to be detracted and increased to reach the frontier, which eases its interpretation, and the DDF is units free in the sense that if we multiply inputs and outputs, as well as their directional vector, by the same vector, then the value of β remains unchanged. This is an interesting property if one rescales inputs and outputs for computational reasons when solving (22). However, if the direction is not in the same units of measurement than inputs and outputs, its interpretation differs and this property does not hold. If the unitary vector is chosen, then the distance function β yields the amount in which inputs and outputs need to be decreased and increased to reach the frontier. Therefore, the relationship between the choice of a directional vector, and the scale properties of the distance function should be considered when choosing a particular specification.

6.2.1.1. Economic-based orientations

The above directional distance vectors can be considered exogenous, since they are chosen by the researcher based on *ad-hoc* criteria. A second possibility that endogenizes the choice is based on the economic behaviour of the firm. When market prices are observed and firms have an economic maximizing behaviour, Zofío et al. (2013) introduce a profit efficiency measure that projects the evaluated firm to the profit maximizing benchmark. The latter proposal can be particularized to the cost minimizing and revenue maximizing cases with the directional input and output distance functions. The method introduced by these authors searches for a directional vector $(-g_x^*, g_y^*)$ that endogenizes the projection of (x_i, y_i) to the profit maximizing benchmark: $\pi(p,w)$, represented by the input–output vector $(x^*, y^*) = \arg\max_{x,y} \{py - wx\}$. The specific reference directional vector is $(g_x^*, g_y^*) \in \mathbb{R}^N \times \mathbb{R}^M \setminus \{0_{N+M}\}$, dropping the negative sign from previous notation—reflecting that inputs are to be reduced and outputs increased, as explained in what follows. The associated profit efficiency measure simultaneously solving the directional vector and identifying the profit maximizing benchmark, can be calculated in the following way:

$$D_T^*(x_{i'}, y_{i'}; p, w) = \max_{\beta, \lambda i, g_x^*, g_y^*} \beta \tag{77}$$

$$
\begin{aligned}
\text{s.t.}\quad & \sum_{i=1}^{\mathrm{I}} \lambda_i\, x_{in} \le x_{i'n} - \beta g^*_{x_n},\ n = 1,\ldots,\mathrm{N},\\
& \sum_{i=1}^{\mathrm{I}} \lambda_i\, y_{im} \ge y_{i'm} + \beta g^*_{y_m},\ m = 1,\ldots,\mathrm{M},\\
& \sum_{m=1}^{M} p_m g^*_{y_m} + \sum_{n=1}^{N} w_n g^*_{x_n} = 1,\\
& \sum_{i=1}^{\mathrm{I}} \lambda_i = 1, \quad \lambda \in R^{\mathrm{I}}_{+},
\end{aligned}
$$

By solving (77), we gain information about firm *i'*s profit inefficiency, the profit maximizing benchmark, and the optimal course that it should follow when planning and adopting profit improving strategies. Particularly, when $D^*_T(x_{i'}, y_{i'}; p, w) > 0$, so the firm is profit inefficient, and in conjunction with the value of the directional distance function (22), we can determine whether the source of the inefficiency is technical, $D_T(x_{i'}, y_{i'}; g_x, g_y) > 0$, or allocative, $D_T(x_{i'}, y_{i'}; g_x, g_y) = 0$. Program (77) departs from (22) in two crucial ways. First, as previously remarked, the directional vector is not preassigned, and therefore (77) searches for it given the price normalization constraint. Secondly, the elements of the directional vector $\left(g^*_x, g^*_y\right)$ could adopt any value, positive and negative, as long as $\left(g^*_x, g^*_y\right) \neq \left(0_N, 0_M\right)$. This means that inputs may be increased and outputs reduced when projecting the evaluated firm to the profit maximizing benchmark. This adds further flexibility to the standard definition (22) that constrains changes in production processes to input reductions and output increases.

Also, we note that the choice of orientation has relevant consequences when measuring overall economic efficiency according to (17), as the profit normalizing condition is a function of the directional vector; i.e., $pg_y + wg_x$. The fact that technical efficiency is measured in the $(-x, y)$ direction results in a normalization of profit by the aggregated value of input and outputs, which has been interpreted as the economic 'size' of the firm −Färe and Grosskopf (2000b), but lacks any straightforward interpretation in an overall economic efficiency context. Therefore, the proposal by Zofío et al. (2013), summarized in (77), normalizing the price constraint to $pg^*_y + wg^*_x = 1$, not only identifies the maximum profit benchmark, but also allows the interpretation of the overall, technical and allocative efficiencies in monetary terms (e.g., dollar valued).

Most importantly, a relevant consequence of this proposal is that it renders the decomposition of overall economic efficiency (15)−(18) redundant when inputs and outputs are fully adjustable at the discretion of managers. Indeed, the main implication of the analysis is the characterization of overall economic efficiency as either technical (wrong engineering practices) or allocative (economic mismanagement when demanding and supplying inputs and outputs). This result derives from the fact that the overall economic efficiency is obtained by identifying the profit efficiency measure along the directional vector $\left(g^*_x, g^*_y\right)$, instead of being simply calculated by

subtracting the observed profit from the maximum profit; which in turn allows determining whether the evaluated firm is on the production frontier or not. Accordingly, the endogenous directional vector $\left(g_x^*, g_y^*\right)$ becomes the cornerstone of the overall evaluation of profit efficiency in the orientation that guarantees maximum profit, and without relying on intermediate steps forced by the subjective choice of the directional vector $\left(g_x, g_y\right)$ in the standard approach (22). From a theoretical and conceptual perspective, our proposal solves the *arbitrary* decomposition of profit efficiency as the relative values of the technical and residual allocative efficiencies depend on the exogenous choice of the directional vector. Finally, from a practical managerial and organizational perspective, if one assumes profit maximizing behaviour on the part of firms, profit inefficient firms will not normally be interested in intermediate projections towards the production frontier that would measure technical efficiency in the exogenous direction $\left(g_x, g_y\right)$. Nevertheless, when some output or inputs are exogenous or non-discretionary, it might not be possible to change the production process towards maximum profit, resulting in overall economic decompositions as in (15)−(18).

6.2.1.2. Data-driven orientations.

When selecting a given orientation, several authors, both in the DEA and SFA, rely on the existing data to identify the most relevant peers. Based in part on the initial contribution by Payrache and Daraio (2012), Daraio and Simar (2016) proposed a method that allows choosing context specific (or local) directions for firms, considering as benchmarks those facing similar conditions, and without assuming any economic behaviour as mentioned in the previous section. These conditions can be associated to the closeness of those benchmark peers to the production (input-output) mix of the evaluated firm, or their share of the same contextual conditions (factors); e.g., benchmarks facing the same non-discretionary inputs and outputs. The method is flexible enough to accommodate a single 'egalitarian' direction for all firms, or individual directions for each one of them, but its strength relies on the fact that the method produces an automatic "peer grouping" of the firms as by-products, depending on their comparable circumstances and external conditions. This last feature is what represents the data-driven approach these authors refer to.

The method is complex in its implementation, as it defines the directional vector in terms of angles in polar coordinates in the multidimensional input-output space, which nevertheless allows these authors to: 1) impose the same direction (angle) using the average of the angles, rather than the average of the observed input and output quantities (so large individual firms will not weigh more in the egalitarian direction); or 2) consider different directions when the contextual factors, represented by a vector *W*, justify their use. How these external factors in *W* influence the direction (angles) is carried out through nonparametric regression analysis of the direction on *W*. It is then a "local" direction determined by its neighbouring (similar) firms, because the proximity is measured according to the variable *W*, and in terms of a bandwidth previously set: $\left|W_i - w\right| \leq ch$, where $c \leq 1$ is a constant scaling the chosen neighbourhood −e.g., $c = 0.5$.

The method proposed by these authors which considers the input dimension of the firm can be summarized in the following algorithm:

1) For all firms, transform each *n*-dimensional and *m*-dimensional input *x* and output *y* vectors into polar coordinates (r_i, θ_i);
2) Perform the polar nonparametric regression for each component θj on W to estimate $\mathrm{E}(\theta^j | W)$. In this step the bandwidth selection is computed by cross-validation. For each direction (angle) and for each data point the local average estimator is obtained: $\hat{\theta}_i^j = \hat{\mathrm{E}}(\theta^j | W_i) = \hat{m}(W_i)$, providing each firm with the angles: $\hat{\theta}_i = (\hat{\theta}_i^1, ..., \hat{\theta}_i^{n-1})$;
3) From the estimated polar coordinates $(r_i, \hat{\theta}_i)$ it is possible to retrace the Cartesian coordinates resulting in the individual directional vector $g_{xi} = \phi(r_1, \hat{\theta}_i)$;
4) Compute the directional distance function for each firm using these directions;
5) Perform the benchmarking analysis for each firm with values (x_i, y_i, w_i) by:
   a) identifying among the firms in the *W*-neighbourhood of w_i, i.e., falling in the interval $\left| W_i^j - w^j \right| \leq 0.5 h^j$, those that are efficient with a calculated directional distance function equal to 0,
   b) select the nearest efficient neighbours by computing the Euclidean distances from the evaluated firm to them, and,
   c) make a radar plot of (x_i, y_i) against the efficient projection and the unit(s) previously identified as benchmarks.

These authors apply their method to different databases either simulated through a suitable DGP from a Cobb-Douglas form or using the original data from Charnes et al. (1981) on education or the banking data used by Simar and Wison (2007). They compare the results following their data driven method, testing the influence of the external factors *W* with those obtained for alternative orientations such as individual specific distances or the egalitarian approach. The method captures the influence of the contextual factors and provides an efficiency measure that takes into account the particularities of the firms being evaluated with respect to their potential benchmark peers. An implementation of the method in a standard software package would be necessary to popularize this potentially useful, but computationally complex method.

6.2.2. SFA framework

In this subsection we discuss the choice of orientation in the SFA approach, and summarize the most recent proposals which aim to endogenize the orientation using economic-based and data-driven criteria.

As mentioned in the introduction of this section, the choice of orientation in the parametric stochastic framework shows its relevance with respect to the "complexity" of the stochastic part of the model. To see this clearly, assume that we want to estimate technology using a stochastic input-oriented distance function $D_I(x_i^*, y_i^*) \cdot e^{v_i} = 1$, where the asterisk stands for efficient units and *v* is the traditional two-sided noise term capturing random shocks. Assume that $u \geq 0$ measures firms'

inefficiency. If we aim to measure inefficient production in terms of overuse of inputs (in this case we assume that $y = y^*$), the frontier model to be estimated can be written as $D_I(x_i e^{-u_i}, y_i) \cdot e^{v_i} = 0$. The linear homogeneity in inputs of this function implies that this equation is equal to:

$$-\ln x_{1i} = \ln D_I\left(\frac{x_i}{x_{1i}}, y_i\right) + v_i - u_i \,. \tag{78}$$

Note that regardless the functional form of the distance function, the random inefficiency term appears in the model as an additive term, and a standard stochastic frontier model appears. Once a distribution for both noise and inefficiency terms is chosen, the parameters in equation (78) can be estimated simply by ML because the *standard* distributional assumptions provide closed forms solutions for the distribution of the composed error term. However, if inefficient production is measured in terms of output reductions (in this case we assume that $x = x^*$), the model to be estimated can be written as $D_I(x_i, y_i e^{u_i}) \cdot e^{v_i} = 0$. The linear homogeneity in inputs of this function yields the following equation:

$$-\ln x_{1i} = \ln D_I\left(\frac{x_i}{x_{1i}}, y_1 e^{u_i}\right) + v_i \tag{79}$$

As customary, if we assume that the distance function has a flexible function form such as the Translog and the firm only produces a single output, the model to be estimated can be written as:

$$\begin{aligned} -\ln x_{1i} = {} & \beta_0 + \sum_{n=2}^{N} \beta_n \ln\left(x_{ni}/x_{1i}\right) + \frac{1}{2}\sum_{n=2}^{N}\sum_{n'=2}^{N} \beta_{nn'} \ln\left(x_{ni}/x_{1i}\right)\ln\left(x_{n'i}/x_{1i}\right) \\ & + \beta_y \ln y_i + \frac{1}{2}\beta_{yy} \ln y_i^2 + \sum_{n=2}^{N} \beta_{ny} \ln\left(x_{ni}/x_{1i}\right)\ln y_i + \varepsilon_i, \end{aligned} \tag{80}$$

where

$$\varepsilon_i = v_i + \left[\beta_y + \sum_{n=2}^{N} \beta_{ny} \ln\left(x_{ni}/x_{1i}\right)\right] u_i + \frac{1}{2}\beta_{yy} u_i^2 \,. \tag{81}$$

It is worth mentioning here that the composed error term ε involves three random terms, v, u and u^2. As Parmeter and Kumbhakar (2014) point out in a cost setting, the presence of the u^2 term in ε makes the derivation of a closed likelihood function impossible. Thus, this precludes using standard maximum likelihood techniques to obtain the parameter estimates. Similar comments can be made if we were to use a directional or generalized−hyperbolic−distance function. In all cases where we have intractable likelihood functions, they can be maximized by simulated maximum likelihood.[68] A final important remark regarding equations (80) and **Error! Reference source not found.** is that the input 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. The same remark obviously can be made for other primal (and dual) representations of firms' technology.

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

It should be also highlighted that the choice of orientation is also relevant for the statistical properties of the regressors. Indeed, once we choose an orientation, the set of frontier variables changes and the endogeneity or exogeneity nature of these variables might also change as well. For instance, while two individual inputs (outputs) might be correlated with the noise (inefficiency) term, their ratio might not if random shocks (inefficient behaviour) have a symmetric effect on both inputs (outputs). In this sense, it is customary in the literature to make endogeneity or exogeneity assumptions on both inputs and outputs based on the economic objectives pursued by the firms. For instance, endogeneity of outputs is often associated with revenue maximization. In this case, the inputs are treated as exogenous variables and the proper orientation to measure firms' inefficiency would be an output-oriented measure. On the other hand, when inputs are treated as endogenous in the cost minimization case, outputs are considered as (demand) predetermined and thus exogenous. In this case, the appropriate measure of technical efficiency is an input-oriented measure. However, if firms pursue maximizing profits, both inputs and outputs should be treated as endogenous variables, and the endogeneity problem remains no matter whether one estimates input or output distance functions. Moreover, Tsionas et al. (2015) also show that the input ratios in an input distance function could be endogenous if allocative errors are correlated with technical inefficiency and/or random productivity shocks. Obviously, a similar conclusion can be obtained for the output ratios in an output distance function.

In Section 6.1 we have summarized several statistical-based strategies to deal with endogeneity issues in stochastic frontier models, such as the use of copulas or ad-hoc reduced form equations for the endogenous variables. We briefly add to that discussion the comment that joint endogeneity of inputs and outputs can be addressed taking advantage of *economic theory,* as advocated by Kumbhakar et al. (2013) and Tsionas et al. (2015). That is, a plausible empirical strategy in the face of overall endogeneity of inputs and outputs would be to augment the input or distance functions with a set of first-order conditions of profit maximization.

6.2.2.1. Data-driven orientations

So far, we have implicitly assumed that the researcher selects a particular orientation before carrying out production and efficiency analyses. The selection is normally based on the features of the industry being examined; e.g., on whether input or outputs are exogenously determined. However, as in its non-parametric DEA counterpart, the input-output orientation issue may also be viewed as a data-driven issue, and thus the decision can be based on performing proper model selection tests. For instance, Orea et al. (2004) find that the model selection approach of Vuong (1989) is a potentially useful tool for identifying the "best" orientation before carrying out such studies. Using a panel-data set on Spanish dairy farms, they fit input, output and hyperbolic-oriented cost frontier models. They performed Vuong tests showing that the input-oriented model is the best among the models estimated; a result that is consistent with the fact that the input - oriented model provides the most credible estimates of scale economies given the structure of the sector.

In the SFA framework, traditional output- and input-oriented models impose a common orientation for all firms over time. The same happens in the paper mentioned above. Kumbhakar et al. (2007) point out this could be a strong assumption in some applications. That is, a given firm could be operating in either regime at any time. For instance, they state that the European railways have changed their strategies from maximizing market share to reducing costs. This suggests that both orientations have played an important role in the European railroad industry and, therefore,

any model used should consider both orientations. These authors treat the input and output distance functions as two latent regimes in a finite mixture model, representing firms' technology by a general stochastic distance function: $0 = \ln D(x, y, \beta) + v \pm u$, where β is a vector of technological parameters, and u is a one-sided random variable representing technical inefficiency whose sign depends on the chosen orientation. The corresponding homogeneity conditions are also imposed. The determination of the efficiency orientation for each firm is addressed by adopting a latent class structure so that the technologies and the probability of being in the input/output oriented inefficiency model are estimated simultaneously by ML. The contribution of firm i to the likelihood is:

$$LF_i = LF_i^I(\gamma_I)\Pi_i(\theta) + LF_i^O(\gamma_O)(1 - \Pi_i(\theta)), \tag{82}$$

where LF^I is the likelihood function of an input distance function model, LF^O is the likelihood function of an output distance function model, $\Pi(\theta)$ is the probability of being in the input-oriented class, and $1\text{-}\Pi(\theta)$ is the probability of being in the output-oriented class. Following Greene (2005) they parameterize the probability of being in the input class $\Pi(\theta)$ as a multinomial logit function that depends on a set of firm-specific variables. The computed posterior probabilities are then used to know whether a particular firm is maximizing output (revenue) or minimizing input use (cost). In essence, the Kumbhakar et al. (2007) model allows the data to sort themselves into the input and output-oriented regimes rather than arbitrarily assuming that all observations obey one or the other at the outset.

6.2.2.2. Endogenous orientations

The latent class model used in Kumbhakar et al. (2007) allows different orientations in an exogenous fashion. There are more probable efficiency measures than others, but the latent class structure of the model does not allow firms to choose the orientation (i.e., the economic objective) they wish to pursue. Therefore, one interesting extension of this model is to endogenize the selection of the orientation of the efficiency measures. This likely can be carried out by adapting one of the models recently introduced in the SFA literature to deal with sample selection problems for this setting. For instance, Greene (2010) and Kumbhakar et al. (2009) use a stochastic frontier sample selection model to take into account the endogeneity of technology choice in estimating the production technology. Later on, Lai (2013) suggests using a threshold stochastic frontier model as the existence of an endogenous threshold variable is analogous to the stochastic frontier sample selection. The key feature of these two models is that production technology is a decision made by the firm itself and thus renders the sample split variable endogenous. The direct consequence of ignoring the endogeneity of the sample split variable is the estimation bias of the production technology, even if the differences in technology (in our case, efficiency orientations) are allowed in the model.

Atkinson and Tsionas (2016) pursue a similar objective using directional distance functions. The typical fixed-direction approach often assumes +1 directions for outputs, and -1 directions for inputs. They argue, however, that since goods (inputs) are produced (demanded) by firms, their relative valuation may not be 1-to-1 for all firms. They generalize the standard (and restricted) models by jointly estimating a quadratic technology-oriented directional distance function, not with directions chosen a priori, but with chosen optimal directions that are consistent

with cost minimization or profit maximization. They first consider the typically-employed quadratic directional distance function of all inputs and outputs as

$$\vec{D}(z,\gamma_g)=0=\sum_{w=1}^{W}\beta_w z_w+\frac{1}{2}\sum_{w=1}^{W}\sum_{w'=1}^{W}\beta_{ww'}z_w z_{w'}-v+u, \tag{83}$$

where $z=(z_1,\ldots,z_W)$ collectively includes all inputs and outputs, γ_g includes all parameters to be estimated, and the proper translation property restrictions have been imposed. They next append price equations (where prices are related to marginal products) for inputs to their directional distance function to obtain a cost-minimization directional distance system and the price equations for all inputs and outputs to obtain a profit-maximization directional distance system:

$$J_w p_w=\left(\beta_w+\sum_{w=1}^{W}\beta_{ww'}z_{w'}\right)p'(J\odot g)-v_w \tag{84}$$

where $J_w=1$ if z_w is an output, and $J_w=-1$ otherwise. Here p denotes the price vector, and v_w is a standard noise component with zero mean, reflecting errors in optimization due to random events beyond the control of the firm. This system of equations is a nonlinear simultaneous equation model where the entire vector z is endogenous. It should be noted that the subscript "g" in (83) indicates that the parameters γ depend explicitly on the direction finally estimated. Therefore, they generalize the dual relationship between the profit function and the technology oriented directional distance function, as established by Chambers (1998), by assuming profit-maximizing behaviour and deriving associated price equations for each input and output. These equations allow identification of directions for each input and output. They estimate the above system using Markov Chain Monte Carlo (MCMC) methods, obtaining estimates of all structural parameters *and* optimal directions. These directions are those that would prevail in the industry if firms were cost minimizers or profit maximizers.

7. Accounting for undesirable production attributes

7.1. Incorporating production risk and stochastic behaviour

Most of the literature measuring firms' production performance lacks an explicit recognition that production takes place under conditions of uncertainty. This may be reasonable in many applications, but not in industries such as agriculture, fishing or banking where production uncertainty is relatively high. In these industries, producers will be concerned about risk properties when they choose input levels and/or they consider the adoption and utilization of new technologies. It is only natural, therefore, that risk considerations be taken into account when evaluating producer production performance. Moreover, as Battese et al. (1997) remark, incorporating production risk into SFA models is of particular relevance because the concept of technical efficiency can be viewed as a measure of the degree of utilization of technologies adopted in the production process. However, standard SFA and DEA efficiency analyses are concerned with estimating non-stochastic behaviour and non-stochastic technologies. Indeed, 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 actually operate. On the other hand, most DEA models and estimation techniques are intended to represent non-stochastic frontiers.

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. For instance, Pope and Chavas (1994) demonstrate that cost minimization cannot be adequately characterized by expected output alone under risk aversion, because the role of risk management in input use can be relevant. This implies that efficient input combinations under production risk can be wrongly labelled as allocative inefficient. 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. In particular, it can lead to the miss-classification of technically and allocative efficient producers as technically inefficient. Finally, uncertainty in demand may also explain inefficiency behaviour in many service industries that require certain endowments of fixed and quasi-fixed assets to satisfy this demand. For instance, Lovell et al. (2009) show that demand uncertainty may affect hospital' costs and that ignoring these effects may lead to biased parameter estimates and misleading inference. Tovar and Wall (2014) show that overcapitalization or infrautilization in the case of Spanish ports can be explained by different demand uncertainty.

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.[69] 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 constructed an econometric model that explicitly accounts for both inefficiency and risk. Following Battese et al. (1997), this author proposed estimating the following SFA model:

$$y_i^t = f(x_i^t, t) + g(x_i^t, t)\,\varepsilon_i^t, \tag{85}$$

where $f(\cdot)$ is the mean function and $g(\cdot)$ is the output risk function, and ε_i^t is a composed error term that includes a noise and an inefficiency term. As we are interested in changes in productivity and welfare over time, we here have added time trend in the production function and a t superscript to all variables. If the variance of the composed random term is normalized to 1, the variance of output is therefore $g(\cdot)$. In this framework, an input is risk-increasing (reducing) (neutral) according to $\partial g(x_i^t)/\partial x_i^t > (<)(=) 0$. Kumbhakar assumed later on that producers maximize the expected utility of anticipated profits, $E[U(\pi)]$. Assuming a single input, the first-order condition of the above problem can be expressed as:

$$\frac{\partial f(x_i^t, t)}{\partial x_i^t} = w - \theta(\cdot)\frac{\partial g(x_i^t, t)}{\partial x_i^t}, \tag{86}$$

[69] We do not summarize here the literature examining firms' performance when input or output prices are uncertain. For a brief revision of this literature using both parametric and non-parametric techniques see Cherchye and Post (2003).

where w is the input price relative to the output price, and $\theta(\cdot)$ is a risk preference function that measures firms risk aversion.[70] This function takes values less than, equal to or higher than zero when producers are risk-averse, risk-neutral or risk-loving, respectively (see Chambers, 1983). 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 deriving an algebraic form of the risk preference function that keeps the model simple for estimation purposes.

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. These authors defined an index of welfare change (WC) in output terms as:

$$WC = \frac{d \ln W}{dt} = \frac{\partial \ln f(x_i^t, t)}{\partial t} - p \cdot r_A \cdot f(x_i^t, t) \frac{\partial \ln g(x_i^t, t)}{\partial t} \cdot m^2, \tag{87}$$

where $r_A(\cdot)=-\theta(\cdot)/\sigma$, is the Arrow-Pratt coefficient of absolute risk aversion, σ is the standard deviation of (normalized) profit, and m is the coefficient of variation of output. This equation suggests that positive technical change ($\partial \ln f/\partial t>0$) increases welfare. In the presence of production risk ($m>0$), increases in the cost of risk reduce welfare provided that producers are risk-averse ($r_A>0$). In particular, if technical change increases (reduces) production risk, i.e., $\partial \ln g/\partial t>(<)0$, producer welfare falls (rises).

As in previous papers focused on production risks, other authors also used simple representations of firms' technology to examine the effect of demand variability on firm costs (profit). See, for instance, Lovell et al. (2009), Rodríguez-Álvarez et al. (2012) and Tovar and Wall (2014). They all assume that firms first choose fixed and quasi-fixed inputs which define service capacity subject to the constraint that the firm wishes to satisfy all but a small proportion of random demand. Once this has been done, in a second stage they choose variable inputs to meet the actual realized demand. Tovar and Wall (2014) argue that if firms cannot adjust inputs after choosing service capacity, then costs will depend not on realized demand or output (y) but rather on service capacity (y^*). In this case the firm incurs costs to produce the capacity to provide a service at a determined level rather than the observed output. Thus, in their case, the cost function facing the port can be expressed as $C=C(y^*,w)$. However, in a more general case some inputs will vary and can be adjusted to meet actual realized demand. In this case, costs will also depend on the realization of demand and actual output (y) should be included as an additional variable in the cost function, that is $C=C(y,y^*,w)$. Gaynor and Anderson (1995) showed, however, that if firms are assumed to have a target service capacity, which is such that the probability that service capacity exceeds demand, is at or above a given target level, and the demand distribution is normal, the service capacity y^* can be replaced with the standard deviation of demand. Thus, the cost function to be estimated can be expressed as $C=C(y,\sigma_y,w)$, where σ_y is the standard deviation of the demand distribution. In this model, realized demand is included to capture the possibility that some inputs can be adjusted after demand is realized. Therefore, the existence of quasi-fixed inputs plus target

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

service capacities requires extending the traditional cost models by including a measure of demand uncertainty.

Many authors examining firms' efficiency in the banking (insurance) industry have followed a similar strategy to account for riskiness and the quality of bank output. Indeed, Mester (1996) pointed out that risk-averse banks tend to fund their loans with a higher ratio of financial capital-to-deposits than risk-neutral banks. Since financial capital is typically more expensive than deposits, this might lead one to wrongly conclude that risk-averse banks are using the wrong input mix, when actually they simply have different risk-preferences than more risk-neutral banks. In order to control for these differences in risk-preferences, this author suggests estimating a cost frontier model with the level of financial capital, that is $C=C(y,w,k)$ where k is the level of financial capital. Other studies such as Altunbas et al. (2001) use equity capital as a control for risk. Altunbas et al. (2000) and Pastor and Serrano (2005) extended the previous cost function in order to investigate the impact of quality factors on banks' cost. In addition to the inclusion of financial capital to control risk, they incorporate loan-loss provisions to control for output quality. In this case, the cost model to be estimated is $C=C(y,\underline{w},k,f)$ where f is a measure of nonperforming loans. They show that if risk and quality factors are not taken into account optimal bank size tends to be overstated. They also show that failure to adequately account for risk can have a significant impact on relative efficiency scores. Pasiouras (2008) deals with the same issues using a non-parametric approach. This author includes nonperforming loans as an additional input in the DEA model to account for credit risk, and found that the inclusion of this variable increased the efficiency scores of the Greek banks.

A common feature of all previous models is that they are 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. For instance, standard models do not separate inefficiency from poor results due to adverse environmental conditions outside the control of the firm. 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.[71] They also show that all the tools of modern production theory, including cost and distance functions, may be applied to state-contingent production technologies.

[71] Rasmussen (2004) distinguishes between different types of inputs according to their influence on production in different states of nature. Thus, state-general inputs are inputs which affect production in some or all states of nature. A state-specific input is one that affects production in only one state of nature, and is therefore a special case of a state-general input. A state-allocable input is one that can affect production in two or more states of nature and which may be allocated (ex-ante) to different states of nature. Rasmussen (2004) refers to such inputs as 'strictly state-allocable' as they only affect production in one state of nature, but notes that there may exist state-allocable inputs which are not strictly state-allocable in the sense that they may be more effective in one state but still have a non-zero influence in other states. From an analytical perspective, state-allocable inputs can be subsumed under state-general or state-specific inputs.

Although Chambers and Quiggin (2000) advocated the use of state-contingent production technologies to represent risky production, empirical application of the state contingent approach has proved difficult for several reasons. First, 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). This creates identification problems and other empirical difficulties in estimating the production technology. For instance, the state-contingent technologies that can be estimated are quite restrictive. It is worth noting that most applications use parametric/econometric techniques. In contrast, Guesmi and Serra (2015) used non-parametric (DEA) techniques to empirically estimate environmental efficiency in a state-contingent framework.

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). In this context, it is assumed that inputs have different marginal effects on outputs depending on the state which occurs, and thus firms may only substitute between state-contingent outputs by choosing different input vectors. To formalize their model, assume a producer uses the vector of N inputs $x=(x_1,\ldots x_N)$ to produce a single stochastic output, y. In this setting, the state of nature is realised after production decisions have been made. Denoting the set of states of nature by $\Omega=(1,\ldots,S)$, then $y=(y_1,\ldots,y_S)$ represents the state-contingent output and y_s represents the amount of output realized in state $s \in \Omega$. As customary, we will label state 1 as "very poor environmental conditions" and state S as "excellent environmental conditions".

The problem is that only one of the S possible output realizations is typically observed. With ex-ante outputs being incompletely observed, this means that neither the cost function $C(w,y_1,\ldots,y_S,t)$ or the distance function $D(y_1,\ldots,y_S,x,t)$ can be estimated. Following Rasmussen (2004), if the state-contingent outputs are independent such that output in a given state does not depend on output in any other state, then production in state s only depends on the input vector x.[72] In this case, the production technology can be described by the set of state-contingent production functions:

$$\ln y_i^t = \alpha_s + f(x_i^t, t, \beta) + v_{is}^t - u_{is}^t, \tag{88}$$

where α_s is a state-varying intercept that allows expected log-output to vary across the states of nature, and v^t_{is} is a normal random variable representing the combined effects of measurement errors and errors arising from the use of approximating functional forms. The standard deviation of this random error is assumed state-dependent. Technically inefficiency will also be expected to differ across states. As noted by O'Donnell and Griffiths (2006), state-contingent frontiers should lead to higher estimates of technical efficiency, as technical inefficiency will be expected to be partly explained by different states of nature; i.e., conventional stochastic frontier models decompose deviations from the frontier into inefficiency and noise whereas state-contingent frontiers decompose deviations from the frontier into inefficiency, noise and risk.

[72] This assumption can be viewed implicitly as a separability restriction on either the above cost or distance functions in that only one (the realized) output is included as regressor, and the other (unrealized) outputs now belonging to the noise term are not correlated with the realized output. That is, this assumption implies that $C=C(w,y_s,t)+\varepsilon$, or $D=D(y_s,x,t)+\varepsilon$ where s is the realized output and $\varepsilon=\varepsilon(y_1,..y_{s-1},y_{s+1},\ldots,y_S,t)$. This separability restriction and the exogeneity assumption of the realized output can be viewed as too strong restrictions in many applications.

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 MLE.

It should be noted that there is an identification (or labelling) problem with the state-contingent latent class model. Assume, for instance, that there are only two different states of nature representing environmental conditions that are unfavourable ('bad' state) and favourable ('good' state) to production. Which class should be labelled as a 'bad' or 'good' state? Our solution to the identification problem is to impose $\alpha_1<\alpha_2$ in equation (88). This labelling restriction ensures that expected log-output increases as environmental conditions improve. Note, however, that the elasticity of expected output in state s with respect to the input in equation (88) is state-invariant. This property may be implausible in some production contexts (e.g., irrigation in rainy and dry seasons). To allow for such a possibility, the slope coefficients in equation (88) must be permitted to vary across states of nature. To solve the identification problem when β is allowed to vary with s=1, O'Donnell and Griffiths (2006) suggest scaling the inputs so that x_i^t=0 at the sample mean.[73] Then the constraint $E(\ln y_i^t \mid x_i^t = 0, s = 1) \leq E(\ln y_i^t \mid x_i^t = 0, s = 2)$ is equivalent to the previous labelling restriction. In this case, however, the constraint is only imposed for the 'representative' firm; i.e., it only can impose a better outcome for the 'good' state locally, and not globally as before. O'Donnell and Griffith (2006) rely on Bayesian estimation to address the identification problem and impose the labelling restriction globally.

Coinciding with previous research that commonly assumes an 'output-cubical' technology, Chavas (2008) proposes a method that allows the researcher to examine substitution possibilities among state-contingent outputs states, and test whether or not the state-contingent technology is 'output-cubical'. Under production uncertainty, the cost function that should be estimated is $C(w,y_1,\ldots,y_S,t)$. In order to make this cost function empirically tractable, this author proposes a method to generate all possible outputs y based on the T observations of the firm. This is done by treating the states as random variables, and making stationarity assumptions on the probability distribution generating these random variables.[74] For example, in the single output case (M = 1), assuming that the states are independently distributed across observations, the regression of output on input use provides a framework to estimate a stochastic production function, where the presence of heteroscedasticity can reflect the effects of input use on the variability of output (e.g., Just and Pope, 1978). The proposed approach has other attractive characteristics. For instance, it does not require a priori risk assessments. In addition, the analysis applies irrespective of risk preferences. To the extent that assessing risk preferences is often difficult, this extends the scope of applications of the proposed method. In contrast, its main limitation 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.

[73] Note that this implies estimating a Translog production function in its approximated form, where the first-order coefficients can be interpreted as elasticities evaluated at the sample mean.

[74] Serra et al. (2010) extended this static cost minimization model and present a dynamic dual model of dynamic decision making under inter-temporal cost minimization in a state-contingent setting and, like Chavas, measure the state-contingent, ex-ante output by simulating an output distribution using the ex-post observations.

7.2. Environmental efficiency

In this section we discuss the different proposals scholars have made from a production theory perspective to model technologies accounting not only for the intended production of firms with a market orientation, but also for undesirable production that, generating negative externalities, should be incorporated into the efficiency analysis.[75] This allows obtaining environmental friendly estimates that do not ignore production that is detrimental to eco-systems, as otherwise one may reach biased conclusions leading to wrong analyses, decision making and policy recommendations.[76]

Färe et al. (1986) were the first to use the output distance function (2) to assess the environmental efficiency of a set of US steam electric plants. However, within an environmental efficiency context, a drawback of the output distance function is that it treats both output subsets symmetrically. As a result, defining radial efficiency measures for equiproportional increases of all outputs—desirable and undesirable—keeps their relative ratios unchanged, and therefore productivity improvements correspond to the classical output to input notion—i.e., the ability to produce more outputs with the same amount of inputs—but not to an environmental productivity notion —i.e., the ability to produce more desirable outputs with *less* undesirable outputs given the same amount of inputs. Therefore, determining relative productivity or efficiency by way of the radial output distance function does not have any implication in terms of environmental efficiency, which remains unchanged, i.e., a business–as–usual strategy. Later on Färe et al. (1989) acknowledged this limitation and measured environmental efficiency relying on the hyperbolic distance function (3), which increases desirable outputs while reducing their undesirable counterparts. From then on, it has been acknowledged that both sets of outputs should be treated differently, crediting firms for their ability to adopt environmentally friendlier benchmarks.

In this section we focus on two of the most important issues that have arguably driven the literature on the subject over the years. On one hand the axiomatic characterization of the technology, and the need to jointly model desirable and undesirable outputs and their physical relation, most notably the existing trade-offs in the form of their—engineering—marginal rates of transformation, represented by shadow prices. On the other, and linked with the general layout of this paper, the possibility of representing the technology through a distance function that, most notably and as explained above, must be capable of treating both sets of outputs asymmetrically. On these grounds, and to structure the presentation, we first present the technology axioms, and later on resort to the directional and generalized distance functions, counterpart to (4) and (5), and

[75] In the literature, 'bad' output is also referred to indistinctly as 'undesirable', 'detrimental', or 'unwanted', corresponding to pollutants, waste, contaminants, etc., which from an economic perspective are produced without the intention to be transacted in markets. By contrast, the 'good' outputs, also referred to as 'desirable' or 'intended', are market oriented, and therefore supplied by the firm with the purpose of maximizing profit. Pittman (1983) initiated the current efficiency and productivity literature accounting undesirable outputs, adopting this term.

[76] Dakpo et al. (2016) discusses, from a broader perspective, the different approaches that have been proposed in the literature to undertake environmental studies. This section corresponds to the environmental performance indicator literature, related to the use distance functions as environmental efficiency measures (scores). For earlier discussions of this literature, including references to its Activity Analysis (DEA) operationalisation—see Tyteca (1996, 1997) and Førsund (2009).

their calculation/estimation by way of DEA and SFA techniques. From there on in Section 8.6 we extend the discussion of environmental efficiency measurement to that of productivity change.

7.2.1. Modelling the production technology with desirable and undesirable outputs

Departing from the characterization of the production possibility set in Section 2, the technology now incorporates undesirable outputs and, therefore it is defined as $T=\left\{(x,y,b): x\in\mathbb{R}_{+}^{N}, y\in\mathbb{R}_{+}^{P}, b\in\mathbb{R}_{+}^{Q}, x \text{ can produce } (y,b)\right\}$, where y and b represent the desirable and undesirable (bad) outputs vectors, respectively. Assuming again the existence of $i = 1,...,I$ firms, the production technology can be equivalently represented, for convenience, by way of the following output correspondence P: $\mathbb{R}_{+}^{N}\rightarrow P(x)\subseteq \mathbb{R}_{+}^{P+Q}$, $P(x)=\left\{(y,b): x \text{ can produce } (y,b)\right\}$. The following axioms are assumed—e.g., Färe et al. (2007): (A1): $0_{P+Q}\in P\left(x\right)$; (A2): $P\left(x\right)$ is compact; (A3): if $x'\geq x$, then $P\left(x\right)\subseteq P\left(x'\right)$; (A4): $(y,b)\in P(x)$ and $0\leq\theta\leq 1$ imply $(\theta y,\theta b)\in P(x)$; (A5): $(y,b)\in P(x)$ and $y'\leq y$ imply $(y',b)\in P(x)$; and (A6): if $(y,b)\in P(x)$ and $b=0$, then $y=0$.

(A1) and (A2) recall the basic regularity conditions that are imposed on the production possibility set. From the perspective of modelling the expanded production technology, (A4) states that a reduction in undesirable output is feasible only if goods are simultaneously reduced, given a fixed level of inputs; i.e., a reduction of the former carries a loss in the latter, and therefore decreasing (disposing of) them is costly−i.e, weak disposability of undesirable outputs; (A5) states that desirable outputs are, on the other hand, strongly (freely) disposable−as inputs by (A3); and (A6) assumes that desirable outputs cannot be produced without waste−null jointness.

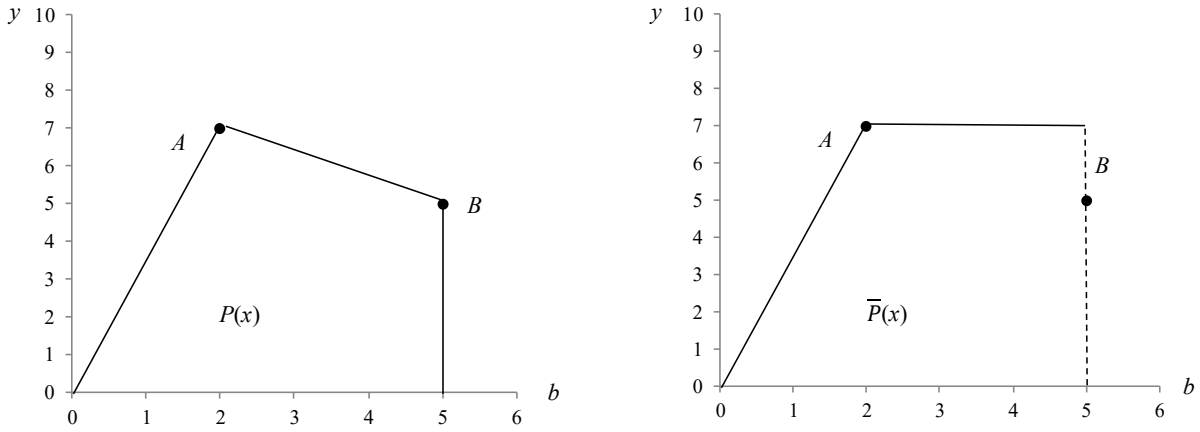
There have been several disputes and controversies in the literature around these axioms characterizing the production technology, and how to model undesirable outputs. The first one corresponds to their consideration as strongly (freely) disposable inputs. An exchange over this issue took place in the *Am. J. Agric. Econ.*, between Hailu and Veeman (2001), Färe and Grosskopf (2003) and Hailu (2003); and later on, in the *Eur. J. Oper. Res.* between Seiford and Zhu (2002), Färe and Grosskop (2004) and Seiford and Zhu (2005). In these exchanges the proposal of treating them as inputs cannot be dissociated from the accompanying strong disposability assumption and the associated DEA production possibility set.[77]

Ultimately, among other reasons, modelling them as strongly disposable inputs solves some problems associated with the previous postulates. We resort to Figure 2 portraying the corresponding DEA characterization of the production possibility set according to the axioms of Chung et al. (1997) to illustrate these problems. First, the production frontier of this set identifies firms belonging to the segment $\overline{AB}$ and the vertical extension as efficient, regardless of their greater undesirable production compared with their truly efficient counterparts in $\overline{0A}$. Consequently, these—wrongly categorized— efficient firms can emerge as reference benchmarks for inefficient firms when taking into consideration the directional and generalized distance functions. Secondly the upward and downward sloping segments associated to the negative and

[77] Seiford and Zhu (2002) propose a data translation method of the undesirable outputs that allows treating them in the DEA framework as if they were inputs, and thereby using the radially oriented model actually reduces them. This transformation however results in a characterization of the technology comparable to that of Hailu and Veeman (2001).

positive shadow prices between the desirable and undesirable outputs, imply that, in the latter case, an efficient firm−or projection to that segment−can reduce undesirable production while increasing desirable production, when only non-positive shadow prices are theoretically acceptable, Lee et al. (2002), Leleu (2013). Thirdly, in a productivity context like that represented by the Malmquist-Luenberger index introduced by Chung et al. (1997)—see Section 8.6 below, Aparicio et al. (2013) show that the weak disposability axiom (WDA) results in an inconsistent interpretation of the technical change component, leading to erroneous conclusions with respect to technological progress or regress, which ultimately plague the productivity index itself.[78]

However, while modelling undesirable outputs as strongly disposable inputs in a DEA context formulation ensures that the model appropriately recognizes that their abatement is costly and is subject to negative shadow prices, it results in an unbounded output production possibility set. This is contrary to the physical laws accounted for by (A2), as an infinite amount of undesirable outputs could be produced by a finite amount of inputs, Podinovski and Kuosmanen (2011).



Figures 3.2.a−b Environmental production sets without and with axiom (A7)

Despite this seemingly irreconcilable trade-off between treating undesirable production as outputs subject to weak disposability with the above setbacks, or handling them as if they were strongly disposable inputs needed to obtain desirable production but violating basic technological assumptions, Aparicio et al. (2013) settle the issue. They consider an additional technological axiom that, departing from the latter option, solves the shortcomings of both proposals and allows for a consistent environmental efficiency and productivity analysis.[79] These authors introduce it as

[78] For these drawbacks associated to the WDA, besides Hailu and Veeman (2001), see Murty et al. (2012) and Leleu (2013) who show that the directional and generalized−hyperbolic− distance functions may generate results that are inconsistent with the Material Balance Equation−and trade-offs between desirable outputs, undesirable outputs, and inputs, as well as incorrect signs for their shadow prices, respectively. Additional evidence related to the diverse disposability characteristics of specific undesirable outputs−some weak, some strongly disposable−considering real technologies has been presented by Yang and Pollit (2010) and Levkoff (2011). Finally, Chen (2014) shows that, regarding the distance functions, the weak disposability assumption is not monotonic in undesirable outputs, and a firm's efficiency can increase (decrease) even if undesirable outputs augment (reduce).

[79] Abiding by the principle of parsimony, these authors searched for a new axiomatic framework that would reconcile both approaches. Such a solution should be compatible with the use of distance functions as suitable characterizations

follows: Given $x \in \mathbb{R}_{+}^{N}$, let $\bar{b}(x): \mathbb{R}_{+}^{N} \to \mathbb{R}_{++}^{Q}$ be a correspondence representing the upper bound for the generation of each undesirable output from the input vector x. Then they introduce the following axiom. (A7): If $(y,b) \in P(x)$ and $b \le b' \le \bar{b}(x)$ it follows that $(y,b') \in P(x)$. In other words, if for a given x the vector (y,b) is feasible, then any vector below the observed upper bound belongs to the production possibility set. Therefore, along with axioms (A1)-(A6) the new axiom that reconciles both approaches establishes that if x can produce outputs (y,b), then it is feasible to produce more contaminants up a certain technological limit, $\bar{b}(x)$, which corresponds to engineering—physical—conversion factors, and can be proxied empirically by the maximum observed amount of undesirable production that is observed. In Figure 2a $\bar{b}(x) = 5$, corresponding to the amount produced by the inefficient firm *B*. The inclusion of the axiom is presented in 3.2b, with the dashed lines highlighting that firms situated in that segment are inefficient.

7.2.2. The directional and generalized distance functions: DEA and SFA formulations

Once the technology has been defined, it can be characterized by distance functions that allow for environmental efficiency improvements by increasing desirable and reducing undesirable production. This restricts the choice to the modified versions of directional and generalized distance functions (4) and (5), which, accounting for undesirable outputs, define as follows:[80]

The environmental DDF: $D_T(x,y,b;g_y,-g_b) = \max\{\beta : (y+\beta g_y, b-\beta g_b) \in \bar{P}(x)\}$, (89)

The environmental GDF: $D_G(x,y,b;\alpha) = \min\{\delta : (y/\delta^{\alpha}, \delta^{1-\alpha} b) \in \bar{P}(x), 0 \le \alpha \le 1\}$. (90)

Both distance functions can be implemented relying on DEA and SFA methods. Regarding the activity analysis (DEA), the production possibility set corresponding to the above (A1)-(A7) axioms is—Aparicio et al. (2017):

$$\bar{P}(x) = \left\{ \begin{array}{l} (y,b) \in \mathbb{R}_{+}^{P} \times \mathbb{R}_{+}^{Q} : \sum_{i=1}^{I} \lambda_i y_{ip} \ge y_p, \;\; p = 1,...,P, \sum_{i=1}^{I} \lambda_i b_{iq} \le b_i, \;\; q = 1,...,Q, \\ \sum_{i=1}^{I} \lambda_i x_{in} \le x_n, \;\; n = 1,...,N, b_q \le \bar{b}_q(x), \;\; q = 1,...,Q, \lambda_i \ge 0, \;\; i = 1,...,I \end{array} \right\}. \quad (91)$$

Considering the directional vector $(g_y, -g_b)$ and α parameter, and following the notation introduced in (22) and (23) allowing for contemporary and "mix-period" distance function that will be recalled in Section 8.6, the mathematical programs associated to the directional and

of the production technology, while from an empirical perspective it should not increase the complexity of the mathematical programming, or result in additional computational burdens. The solution based on the new postulate requires minimal changes to the widely accepted axioms (A1)-(A6), as well as for its DEA implementation, which is presented in Aparicio et al. (2017).

[80] The environmental distance functions are defined in terms of the output production possibility set $\bar{P}(x)$, effectively rendering them output distance functions; However, they could be enhanced to allow for input reductions. Again, it remains to be explored how this change in the production plans could conflict with other approaches such as the material balance principle.

generalized distance functions evaluating firm *i'* in period *t*=0, 1, with respect to the frontier of technology $\overline{P}^s$ in period *s*=0, 1, are:

Environmental *DDF*

$$
\begin{aligned}
&\mathrm{D}_T^s(x_{i'}^t, y_{i'}^t, b_{i'}^t; g_y, -g_b) = \\
&= \min_{\beta,\lambda_i} \left\{ \beta : (y_{i'}^t + \beta g_y, b_{i'}^t - \beta g_b) \in \overline{P}^s(x) \right\} \\
&\text{s.t.} \quad \sum_{i=1}^{I} \lambda_i y_{ip}^s \geq y_{i'p}^t + \beta y_{i'p}^t, \ p = 1,...,P, \\
&\qquad \sum_{i=1}^{I} \lambda_i b_{iq}^s \leq b_{i'q}^t - \beta b_{i'q}^t, \ q = 1,...,Q, \\
&\qquad \sum_{i=1}^{I} \lambda_i x_{in}^s \leq x_{i'n}^t, \ n = 1,...,N \\
&\qquad b_{i'q}^t - \beta b_{i'q}^t \leq \overline{b}_q^s \left(x_{i'}^t \right), \ q = 1,...,Q, \\
&\qquad \lambda_i \geq 0.
\end{aligned}
\tag{92}
$$

Environmental *GDF*

$$
\begin{aligned}
&\mathrm{D}_G^s \left(x_{i'}^t, y_{i'}^t, b_{i'}^t; \alpha \right) = \\
&= \min_{\delta,\lambda_i} \left\{ \delta : (y_{i'}^t / \delta^{\alpha}, \delta^{1-\alpha} b_{i'}^t) \in \overline{P}^s(x) \right\} \\
&\text{s.t.} \quad \sum_{i=1}^{\mathrm{I}} \lambda_i \, y_{ip}^s \geq y_{i'p}^t, / \delta^{\alpha}, \quad p = 1,...,P \\
&\qquad \sum_{i=1}^{\mathrm{I}} \lambda_i \, b_{iq}^s \leq \delta^{1-\alpha} \, b_{i'q}^t, \quad q = 1,...,Q, \\
&\qquad \sum_{i=1}^{\mathrm{I}} \lambda_i \, x_{in}^s \leq x_{i'n}^t, \quad n = 1,...,N, \\
&\qquad b_{i'q}^t - \beta b_{i'q}^t \leq \overline{b}_q^s \left(x_{i'}^t \right), \ q = 1,...,Q, \\
&\qquad \lambda_i \geq 0.
\end{aligned}
\tag{93}
$$

These formulations can be seen as a bridge between the two aforementioned approaches in the literature for dealing with desirable and undesirable outputs. Indeed, they force undesirable output projections to be greater than or equal to the benchmark frontier combination—adopting the rationale underlying input modelling, but upper bounding the feasible values. This prevents a finite amount of inputs from producing infinite amounts of undesirable production.

Alternatively, one can resort to SFA in order to estimate the environmental directional and generalized distance functions. As anticipated in Section 3.2, in the parametric case associated with flexible functional forms specifications, whether or not the estimation complies with the technological axioms could be checked locally at the sample mean or by considering individual data−e.g., as is the case with the shadow prices between desirable and undesirable outputs corresponding to their monotonicity conditions.

Considering first the directional distance function, Färe et al. (2005) rely on the quadratic specification and set the directional vector (g_y, -g_b) to (1, −1), which eases the interpretation of the inefficiency values and allows for their aggregation at the industry level as discussed in Section 6.2. In the environmental context, the counterpart to equation (92), corresponds to the following specification:

$$
\begin{aligned}
D_T(x_i, y_i, b_i; 1, -1) = \alpha_0 &+ \sum_{n=1}^{N} \alpha_n x_{ni} + \sum_{p=1}^{P} \chi_p y_{pi} + \sum_{q=1}^{Q} \iota_q b_{qi} + \\
&+ 1/2 \sum_{n=1}^{N} \sum_{n'=1}^{N} \alpha_{nn'} x_{ni} x_{n'i} + 1/2 \sum_{p=1}^{P} \sum_{p'=1}^{P} \chi_{pp'} y_{pi} y_{p'i} + 1/2 \sum_{q=1}^{Q} \sum_{q=1}^{Q} \iota_{qq'} b_{qi} b_{q'i}
\end{aligned}
\tag{94}
$$

With symmetry between cross parameters: $\alpha_{nn'} = \alpha_{m'n}, \chi_{pp'} = \chi_{p'pi}, \iota_{qq'} = \iota_{q'q}, \forall i, j$. For the translation property introduced in the second section to hold: $D_T(x, y - \lambda, b + \lambda; 1, -1) = D_T(x, y, b; 1, -1) - \lambda, \lambda \in \mathbb{R}_+$, the required parameter restrictions are: $\sum_{p=1}^{P} \chi_p - \sum_{q=1}^{Q} \iota_q = -1$,

$\sum_{p=1}^{P}\chi_{pp'}g_{yp}=0,\ p=1,...,P,\qquad \sum_{p'=1}^{P}\chi_{pp'}g_{yp'}=0,\ p'=1,...,P,\qquad \sum_{q=1}^{Q}\iota_{qq'}g_{bq}=0,\ q=1,...,Q,$

$\sum_{q'=1}^{Q}\iota_{qq'}g_{bq'}=0,\ q'=1,...,Q,\qquad \sum_{n=1}^{N}\mu_{np}-\sum_{n=1}^{N}\nu_{nq}=0,\ n=1,...,N$ and

$\sum_{p=1}^{P}o_{pq}g_{yp}=\sum_{q=1}^{Q}o_{pq}g_{bq}=0,\ p=1,...,P,\ q=1,...,Q.$

Regarding the generalized distance function, Cuesta et al. (2009) show that the Translog specification is appropriate for the imposition of the required almost homogeneity property−Aczel, (1966; Chs. 5, 7), Lau (1972): $D_G(\lambda^{k_1}x,\lambda^{k_2}y,\lambda^{k_3}b;\alpha)=\lambda^{k_4}D_G(x,y,b;\alpha),\ \forall\lambda>0$. For the specific case that leaves inputs unchanged, while increasing desirable outputs and reducing undesirable outputs−equation (90), these conditions correspond to k = (0, 1, −1, 1). Also, given the translog specification of the generalized distance function under α = 0.5: [81]

$$\begin{aligned}1/2\ln D_G(x_i,y_i,b_i;\,0.5)=\alpha_0+\sum_{n=1}^{N}\alpha_n\ln x_{ni}+\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{nn'}\ln x_{ni}\ln x_{n'i}+\sum_{p=1}^{P}\chi_p\ln y_{pi}+\\ \frac{1}{2}\sum_{p=1}^{P}\sum_{p'=1}^{P}\chi_{pp'}\ln y_{pi}\ln y_{pi}+\sum_{q=1}^{Q}\iota_q\ln b_{qi}+\frac{1}{2}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\iota_{qq'}\ln b_{qi}\ln b_{q'i}+\\ \sum_{n=1}^{N}\sum_{p=1}^{P}\mu_{pq}\ln x_{ni}\ln y_{pi}+\sum_{n=1}^{N}\sum_{q=1}^{Q}\nu_{nq}\ln x_{ni}\ln b_{qi}+\sum_{p=1}^{P}\sum_{q=1}^{Q}o_{pq}\ln y_{pi}\ln b_{qi},\end{aligned}\tag{95}$$

the specific almost homogeneity can be imposed through the following restrictions: $\sum_{p=1}^{P}\chi_p-\sum_{q=1}^{Q}\iota_q=1,\ \ \sum_{p=1}^{P}\chi_{pp'}-\sum_{p=1}^{P}o_{pq}=0,\ \ p=1,...,P,\ \ \sum_{m=1}^{M}\mu_{mp}-\sum_{m=1}^{M}\nu_{mq}=0,\ m=1,...,M$, and $\sum_{q=1}^{Q}o_{mq}-\sum_{q=1}^{Q}\iota_{qq'}=0,\ \ q=1,...,Q$. Cuesta et al. (2009) discuss the methodology which leads to the set of restrictions necessary for imposing the homogeneity condition k corresponding to alternative directions, e.g., including input reductions, or increasing desirable outputs only.[82] It is possible to impose this set of restrictions on the translog distance function by modifying the approach introduced by Lovell et al. (1994). For the above case corresponding to the almost homogeneity condition k = (0, 1, −1, 1), and choosing the last desirable output for normalizing purposes, $\mu = 1/y_P$, this corresponds to $D_G(x,y/y_P,by_P;\,0.5)=D_G(x,y,v;\,0.5)/y_P$, and (95) can be respecified for estimation purposes by normalizing the desirable and undesirable outputs accordingly:

[81] This particular specification with α = 0.5 corresponds to the hyperbolic distance function, which increases desirable outputs and reduces undesirable outputs by the same factor. In this case the following relationship between the hyperbolic and generalized distance function is verified: $D_H(x,y,b) = D_G(x,y,b;\alpha=0.5)^{1/2}$; i.e., $\ln D_H(x,y,b) = 1/2\ln D_G(x,y,b;\alpha=0.5)$ under the logarithmic transformation necessary to define the translog specification (95).

[82] The enhanced environmental hyperbolic distance function including input reductions is almost homogeneous of degrees k = (-1, 1, -1, 1), while the output distance function, is almost homogeneous of degrees k = (0, 1, 0, 1).

$$1/2\ln\left(\frac{D_G(x_i,y_i,b_i;\,0.5)}{y_{Pi}}\right)=\alpha_0+\sum_{n=1}^{N}\alpha_n\ln x_{ni}+\frac{1}{2}\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{nn'i}\ln x_{ni}\ln x_{n'i}+\sum_{p=1}^{P-1}\chi_p\ln y^*_{pi}+$$
$$\frac{1}{2}\sum_{p=1}^{P-1}\sum_{p'=1}^{P-1}\chi_{pp'}\ln y^*_{pi}\ln y^*_{p'i}+\sum_{q=1}^{Q}\iota_q\ln b^*_{qi}+\frac{1}{2}\sum_{q=1}^{Q}\sum_{q'=1}^{Q}\iota_{qq'}\ln y^*_{qi}\ln b^*_{q'i}+ \tag{96}$$
$$\sum_{n=1}^{N}\sum_{p=1}^{P-1}\mu_{pq}\ln x_{ni}\ln y^*_{pi}+\sum_{n=1}^{N}\sum_{q=1}^{Q}\nu_{nq}\ln x_{ni}\ln b^*_{qi}+\sum_{p=1}^{P-1}\sum_{q=1}^{Q}o_{pq}\ln y^*_{pi}\ln b^*_{qi},$$

where $y^*_{pi}=y_{pi}/y_{Pi}$ and $b^*_{qi}=b_{qi}y_{Pi}$.[83]

Both equations (94) and (96) can be estimated with suitable frontier methods by adding the composed error term $e_i=v_i-u_i$, so as to obtain the individual conditional distribution of the one sided error, representing environmental efficiency as presented in Section 3.2. Afterwards, making use of the derivative properties of the flexible distance functions, technological characteristics in the form of returns to scale, substitutability and shadow-pricing can be determined from the parameter estimates.

7.2.3. Eco-efficiency

The concept of eco-efficiency, complementing the environmental efficiency analysis just presented, is becoming a popular tool to capture economic and environmental aspects of production; e.g., when agricultural activity generates adverse environmental impacts (see Picazo-Tadeo et al., 2011). The measurement of eco-efficiency in a frontier context compares economic value added with aggregate measures of the environmental impacts generated by the production process. The literature to date has used the DEA approach exclusively to measure producers' eco-efficiency. A remarkable exception is Orea and Wall (2017) that have recently showed that it can also be estimated using a SFA approach.

7.2.3.1. The DEA eco-efficiency model

Kuosmanen and Kortelainen (2005) related the concept of eco-efficiency, defined as a ratio of economic—vale added—activity to environmental damage, to frontier analysis. They proposed a pressure-generating technology set: $PGT=\{(v,\ p)\in\mathbb{R}\times\mathbb{R}^N : v \text{ can be generated by } p\}$, which describes all the feasible combinations of economic value, *v*, and environmental pressures or pollutants, *p*—these authors discuss the properties of the technology set. Environmental damage, $D(p)$ is measured by aggregating the *N* environmental pressures $(p_1,\ldots,p_N)$ associated with the production activity.

For producer i, individual eco-efficiency scores can be expressed as:

[83] For the normalizing output y_{Pi} the ratio y^*_{Pi} is equal to one, and all terms involving the normalizing output are null, explaining why the summations involving y^*_{pi} are over $P-1$. It is straightforward to verify that the translog generalized distance function satisfies the desirable properties.

$$EEF_i = \frac{\text{Economic value added}}{\text{Environmental pressure}} = \frac{VA_i}{D_i(p)}. \tag{97}$$

Here, $D_i(p)$ is a function that aggregates the environmental pressures into a single indicator through a linear weighted average of the individual environmental pressures:

$$D_i(p) = w_1 p_{1i} + w_2 p_{2i} + \ldots + w_N p_{Ni}, \tag{98}$$

where w_k is the weight assigned to environmental pressure p_k. As a non-subjective weighting method, Kuosmanen and Kortelainen (2005) resort to Data Envelopment Analysis, characterizing the *PGT* in similar terms to (19). Picazo-Tadeo et al. (2012) propose using the directional distance function (4) to measure the eco-efficiency, adapting it to the current context: i.e., $D_T(v, p; -g_p, g_v) = \max\{\beta : (p - \beta g_p, v + \beta g_v) \in PGT\}$, $g = (-g_p, g_v) \setminus \{0_{N+1}\}$. Note however that it could be alternatively measured by way of the generalized distance function: $D_G(p, v; \alpha) = \min\{\delta : (\delta^{1-\alpha} p, v / \delta^{\alpha}) \in PGT\}$, $0 \le \alpha \le 1$. For firm *i'*, both distance functions can be calculated resorting to the following programmes counterpart to (22) and (23):

Directional Eco-efficiency, *DDF*

$$\begin{aligned} & D_T(p_{i'}, v_{i'}; -g_p, g_v) = \\ & = \max_{\beta, \lambda_i} \{\beta : (p_{i'} - \beta g_p, v_{i'} + \beta g_v) \in PGT\} \\ & \text{s.t.} \quad \sum_{i=1}^{I} \lambda_i \, p_{in} \le p_{i'n} - \beta g_{p_k}, \; N = 1, \ldots, N, \\ & \qquad \sum_{i=1}^{I} \lambda_i \, v_i^s \ge v_{i'}^t + \beta g_v, \\ & \qquad \lambda \in \mathbb{R}_+^I. \end{aligned} \tag{99}$$

Generalized Eco-efficiency, *GDF*

$$\begin{aligned} & \hat{D}_G(p_{i'}, y_{i'}; \alpha) = \\ & = \min_{\delta, \lambda_i} \{\delta : (p_{i'} \delta^{1-\alpha}, v_{i'}^t / \delta^{\alpha}) \in PGT\} \\ & \text{s.t.} \quad \sum_{i=1}^{I} \lambda_i \, p_{in}^s \le \delta^{1-\alpha} \, x_{i'n}^p, \quad n = 1, \ldots, N, \\ & \qquad \sum_{i=1}^{I} \lambda_i \, v_i \ge v_{i'} / \delta^{\alpha}, \\ & \qquad \lambda \in \mathbb{R}_+^I. \end{aligned} \tag{100}$$

Depending on the choice of directional vector, different measures of eco-efficiency ore obtained. Among these, the 'input' oriented measure proposed by Kuosmanen and Kortelainen (2005) and corresponding to $(-g_p, g_v) = (-p, 0)$ and α=0; the 'output' oriented measure with $(-g_p, g_v) = (0, v)$ and α=1; and other intermediate directions increasing value added and reducing environmental pressures—which require further justification as discussed in Section 6.2, including those that assign different directions for each observation. The eco-efficiency literature bypasses the axiomatic debate on how to characterize the production technology presented in the previous section by ignoring the inputs' side, along with constant returns to scale. Indeed, programmes (99) and (100) do not account for input variables—although environmental pressure variables are treated as inputs, which along with the constant returns to scale assumption would imply that, at the eco-efficient frontier, environmental pressures can be decreased by a proportionate downward

scaling of economic—value added—activity, which is closely related to the weak disposability property—Kuosmanen (2005).[84]

7.2.3.2. The SFA eco-efficiency model

When specified in the form of the 'input' or 'output' oriented eco-efficiency, problems (99) and (100) are equivalent to the DEA efficiency measure in ratio—multiplier—form suggested by Kuosmanen and Kortelainen (2005):

$$\begin{aligned} &\max_{w_{ni}} EEF_{i'} = \frac{VA_{i'}}{\sum_{n=1}^{N} w_{ni'} p_{ni'}} \\ &\text{s.t.} \quad \frac{VA_i}{\sum_{n=1}^{N} w_{ni} p_{ni}} \le 1, \quad I = 1,\ldots,I, \\ &\qquad w_{ni} \ge 0, \qquad\qquad n = 1,\ldots,N. \end{aligned} \tag{101}$$

These constraints force weights to be non-negative and eco-efficiency scores to adopt values between zero and one, that is:

$$EEF_i = \frac{VA_i}{\sum_{n=1}^{N} w_{ni} p_{ni}} \le 1, \quad \forall i = 1,\ldots,I. \tag{102}$$

Equation (102) is useful for deriving a stochastic frontier eco-efficiency model. In the parametric setting, the coefficients on the environmental pressures are parameters to be estimated, representing marginal contributions to value added. We can impose non-negativity by reparameterizing them as $w_n = e^{\beta_n}$. Taking logs in (102),

$$\ln EFF_i = \ln\left(\frac{VA_i}{\sum_{n=1}^{N} e^{\beta_n} \cdot p_{ni}} \right) \le 0, \tag{103}$$

which can be rewritten as:

$$\ln VA_i = \ln\left(\sum_{n=1}^{N} e^{\beta_n} \cdot p_{ni} \right) - u_i, \tag{104}$$

where $u_i = -\ln EFF_i \ge 0$ can be interpreted as a non-negative random term capturing firm *i*'s eco-inefficiency.

To incorporate the effects of random shocks on economic value we extend the model in (104) by adding a symmetric random noise term, v_i, and a non-zero intercept θ:

[84] Naturally, constant returns to scale implies that "*... the size of the firm or production activity…does not matter in this problem (the assessment of eco-efficiency); we are only interested about the ratio of the value added to the environmental pressure. In DEA literature* (this) *is interpreted as a constant returns to scale model*", Kortelainen and Kuosmanen (2004, p. 14).

$$\ln VA_i = \theta + \ln\left(\sum_{n=1}^{N} e^{\beta_n} \cdot p_{ni}\right) + v_i - u_i \,. \tag{105}$$

Here, deviations from the frontier due to random noise are incorporated along with eco-inefficiency. The non-zero intercept ensures unbiased parameter estimates in the event that components of random noise have a level effect on firms' economic value.

The error term $\varepsilon_i = v_i - u_i$ in (105) is composed of two independent parts. φ_i is a two-sided random noise term, often assumed to be normally distributed with zero mean and constant standard deviation, i.e., $\sigma_\varphi = e^{\gamma}$. The parameter u_i is a one-sided error term capturing underlying eco-inefficiency. As usual, following Aigner et al. (1977) this is often assumed to follow a half-normal distribution, which is the truncation (at zero) of a normally-distributed random variable with mean zero and standard deviation $\sigma_u = e^{\delta}$. Under these distributional assumptions the density function of the composed error term $\varepsilon_i = v_i - u_i$ in (105) is the same as the density function of a standard normal-half normal frontier model. Following Kumbhakar and Lovell, (2000, p.77), the log likelihood function for a sample of *I* producers can then be written as:

$$\ln L(\theta,\beta,\gamma,\delta) = -\frac{I}{2}\ln\left[\sigma_v^2 + \sigma_u^2\right] + \sum_{i=1}^{I} \ln\Phi\left[-\frac{\varepsilon_i(\theta,\beta)\cdot\sigma_u/\sigma_v}{\left(\sigma_v^2+\sigma_u^2\right)^{1/2}}\right] - \frac{1}{2\left(\sigma_v^2+\sigma_u^2\right)}\sum_{i=1}^{I}\varepsilon_i(\theta,\beta)^2, \tag{106}$$

where $\beta = (\beta_1,\ldots,\beta_K)$, and $\varepsilon_i(\theta,\beta) = \ln VA_i - \theta - \ln\left(\sum_{n=1}^{N} e^{\beta_n} p_{ni}\right)$.

The likelihood function can be maximized with respect to $(\theta,\beta,\gamma,\delta)$ to obtain consistent estimates of all parameters of the eco-efficiency model. Also, given the simplicity of the error term it is possible to rely on Jondrow *et al.* (1982) and use the conditional expectation $E(u_i \mid \varepsilon_i)$ to estimate the asymmetric random term u_i and compute the firm's eco-efficiency score as $\exp(-u_i)$.

In a traditional SFA production model, $\varepsilon_i(\theta,\beta)$ is a simple linear function of the parameters to be estimated whereas in the eco-efficiency frontier model it is a non-linear function of the β parameters. It should also be noted that the SFA model based on (106) can accommodate heteroskedastic inefficiency and noise terms simply by modelling the variances of σ_u and σ_v as functions of some exogenous variables.

As usual, and compared to the DEA eco-efficiency model, the SFA approach attenuates the effects on eco-efficiency scores of outliers and measurement errors in the data. Moreover, the 'technology' in the eco-efficiency model is a simple index that aggregates all environmental pressures into a unique value. Hence the parametric specification of a functional form is not as problematic as it might be in the production setting of the previous section where possible multiple inputs and desirable and undesirable outputs exist. Also, very few parameters need to be estimated, so like DEA, the SFA model can potentially be implemented even when the number of observations is relatively small.

Finally, note that the estimated β parameters have an interesting interpretation in the parametric model. Eco-efficiency is constant and equal to 1 along the eco-efficiency frontier. Differentiating (97) and using the reparameterisation of the pressure weights in (103) we obtain:

$$\frac{\partial VA_i / \partial p_n}{\partial VA_i / \partial p_{n'}} = \frac{e^{\beta_n}}{e^{\beta_{n'}}}. \tag{107}$$

Once the β parameters have been estimated, e^{β_n} represents the marginal contribution of pressure p_n to firm *i*'s economic value; i.e., it is the monetary loss in economic value if pressure p_n were reduced by one unit. As expression (107) represents the *marginal rate of technical substitution of environmental pressures*, it provides valuable information on the possibilities for substitution between pressures as well as elucidating the possible consequences for firms resulting from legislation requiring reductions in individual pressures.

8. Productivity

8.1. Malmquist index and Luenberger indicator

Based on the distance functions (1)−(5), if panel data is available for several periods, $t = 1,\ldots,T$, it is possible to calculate the change in a firm's productivity by way of the multiplicative Malmquist Index and additive Luenberger Indicator. Ultimately, as these two definitions correspond to quantity measures comprising cost, revenue, profitability and profit change, the choice of distance function is capital, not only when measuring productivity change but also if a broader time-series economic analysis is being contemplated.

We depart from the multiplicative Malmquist productivity index, MPI, that was introduced earlier in the literature than its Luenberger additive counterpart. Caves et al. (1982) theoretically introduced a version of the index which although not having the proportionality property, was later on popularized by Färe et al. (1989, 1994), who assuming constant returns to scale rendered it applicable by way of Data Envelopment Analysis techniques. Following Balk (2001), productivity change for a firm observed in the base period ($t = 0$) and comparison period ($t = 1$)−represented by (x_i^0, y_i^0) and (x_i^1, y_i^1), respectively−is measured by some positive, finite function $F(x_i^1, y_i^1, x_i^0, y_i^0)$. This function should be nonincreasing in x_i^1, nondecreasing in y_i^1, nondecreasing in x_i^0, and nonincreasing in y_i^0. Most importantly, it must exhibit proportionality in input and outputs quantities, implying that: $F(\lambda x_i^0, \mu y_i^0, x_i^0, y_i^0) = \mu / \lambda$, $\lambda, \mu > 0$. Considering these desirable properties, the Malmquist productivity index based on the generalized distance function can be conveniently defined as:

$$\hat{M}_G^s(x_i^1, y_i^1, x_i^0, y_i^0; \alpha) = \frac{\hat{D}_G^s\left(x_i^1, y_i^1; \alpha\right)}{\hat{D}_G^s\left(x_i^0, y_i^0; \alpha\right)}, \tag{108}$$

where the MPI is the ratio of two generalized distance functions (5) normally defined under the base (s=0) or comparison period (s=1) virtual (cone) technology characterized by constant returns to scale, CRS. CRS is a prerequisite for the proportionality property to be satisfied. For the MPI

to exhibit this property, the generalized distance function, which is almost homogenous of degree (α–1), α and 1 in x and y—Chavas and Cox (1999; p. 300), readily satisfies it under CRS, with $\hat{D}_G^s\left(x_i^t, y_i^t; \alpha\right)$, $t = 0, 1$, being independent of α. Taking as reference the base period, the numerator in (108) corresponds to $\hat{D}_G^0\left(x_i^1, y_i^1; \alpha\right)$. This distance function evaluates the efficiency (or productivity level) of the firm observed in the comparison period (x_i^1, y_i^1) with respect the technology in the base period $\hat{T}^0$. Recalling program (23), it is calculated considering $s = 0$ as the reference technology in the objective function and the left hand side of the restrictions, and $h = 1$ for the evaluated firm. On the contrary, the counterpart distance function $\hat{D}_G^1\left(x_i^0, y_i^0; \alpha\right)$, necessary to calculate the comparison period ($s = 1$) Malmquist productivity index in (108) and measure technological change in (109) below, requires reversing the time superscripts in (23).

Balk and Zofío (2017) discuss alternative ways to decompose (108) considering the output ($\alpha = 1$) and input ($\alpha = 0$) orientations, and present the methods to identify the relevant sources contributing to productivity change. Here we present the particular decomposition for the output oriented index and the base period reference technology; i.e., $\alpha = 1$ and s=0:

$$\begin{aligned}\hat{M}_O^0(x_i^1, y_i^1, x_i^0, y_i^0) &= \frac{\hat{D}_O^0\left(x_i^1, y_i^1\right)}{\hat{D}_O^0\left(x_i^0, y_i^0\right)} = \frac{D_O^1\left(x_i^1, y_i^1\right)}{D_O^0\left(x_i^0, y_i^0\right)} \cdot \left[\frac{\hat{D}_O^0\left(\lambda x_i^0, y_i^0\right)}{D_O^0\left(\lambda x_i^0, y_i^0\right)} \cdot \frac{D_O^0\left(x_i^0, y_i^0\right)}{\hat{D}_O^0\left(x_i^0, y_i^0\right)}\right] \cdot \\ &\left[\frac{\hat{D}_O^0\left(x_i^1, y_i^0\right)}{D_O^0\left(x_i^1, y_i^0\right)} \cdot \frac{D_O^0\left(\lambda x_i^0, y_i^0\right)}{\hat{D}_O^0\left(\lambda x_i^0, y_i^0\right)}\right] \cdot \left[\frac{\hat{D}_O^0\left(x_i^1, y_i^1\right)}{D_O^0\left(x_i^1, y_i^1\right)} \cdot \frac{D_O^0\left(x_i^1, y_i^0\right)}{\hat{D}_O^0\left(x_i^1, y_i^0\right)}\right] \cdot \frac{D_O^0\left(x_i^1, y_i^1\right)}{D_O^1\left(x_i^1, y_i^1\right)} \\ &= EC_O(x_i^1, y_i^1, x_i^0, y_i^0) \cdot SEC_O^0(\lambda x_i^0, x_i^0, y_i^0) \cdot \\ &SEC_O^0(x_i^1, \lambda x_i^0, y_i^0) \cdot OME^0(x_i^1, y_i^1, y_i^0) \cdot TC_O^{1,0}(x_i^1, y_i^1).\end{aligned} \tag{109}$$

This MPI decomposed into the following terms: i) technical efficiency change: $EC_O(x_i^1, y_i^1, x_i^0, y_i^0)$, measuring the change in the distance of the evaluated firm from the production frontiers in both periods, ii) a radial scale effect: $SEC_O^0(\lambda x_i^0, x_i^0, y_i^0)$, measuring radial scale effects —since $SEC_O^0(\lambda x_i^0, x_i^0, y_i^0) = SEC_O^0(\lambda x_i^0, x_i^0, \mu y_i^0)$, for $\mu > 0$; iii) an input mix effect: $SEC_O^0(x_i^1, \lambda x_i^0, y_i^0)$, iv) an output mix effect: $OME^0(x_i^1, y_i^1, y_i^0)$, and v) technological change: $TC_O^{1,0}(x_i^1, y_i^1)$, measuring the change in the production frontier from the perspective of (x_i^1, y_i^1). Values greater (smaller) than one for any of these terms reflect an incremental (detrimental) contribution to productivity change of each one of these terms.[85]

These five factors are indeed independent, and therefore do not combine different concepts. First, when the firm is technically efficient in both periods, then $EC_O(x_i^1, y_i^1, x_i^0, y_i^0) = 1$. Second, in the absence of technical change: $T^0 = T^1$, and $TC_O^{1,0}(x_i^1, y_i^1) = 1$. Third, if $x_i^1 = \lambda x_i^0$, $\lambda > 0$, the

[85] Merging the radial scale effect and the input mix effect one obtains the decomposition proposed by Balk (2001), while merging the radial scale effect, the input mix effect, and the output mix effect results in that proposed by Ray and Desli (1997).

input mix disappears: $SEC_O^0(x_i^1, \lambda x_i^0, y_i^0) = 1$. Fourth, if $y_i^1 = \mu y_i^0$, $\mu > 0$, the same happens with the output mix effect: $OME^0(x_i^1, y_i^1, y_i^0) = 1$. Therefore, the only remaining term corresponds to the radial scale effect $SEC_O^0(\lambda x_i^0, x_i^0, y_i^0)$. Relying on the homogeneity properties of the distance functions, the proportionality property associated to radial scale effects is verified:

$$SEC_O^0(\lambda x_i^0, x_i^0, y_i^0) = \frac{1}{\lambda D_O^0\left(\lambda x_i^0, y_i^0\right)} = \frac{\mu}{\lambda D_O^0\left(\lambda x_i^0, \mu y_i^0\right)} = \frac{\mu}{\lambda D_O^1\left(x_i^1, y_i^1\right)} = \frac{\mu}{\lambda}. \tag{110}$$

While the value of the Malmquist productivity index is independent of the orientation, it is not the case for the different terms in which it decomposes, and therefore the choice of orientation is not neutral. This supports the choice of a balanced orientation when calculating the MPI, as when α = 0.5—a hyperbolic orientation, and gave rise to alternative productivity indices accounting for the output an input orientations. The latter include the family of Moortseen-Bjurek productivity indices, making use of the input and output distance functions (1) and (2), see Bjurek (1996). Moreover, as the value of the MPI depends on the reference period, the choice of reference period is not neutral either, and the geometric mean yields a value that weights both periods equally:

$$\hat{M}_G^{0,1}(x_i^0, y_i^0, x_i^1, y_i^1; \alpha) = \left[\frac{\hat{D}_G^0\left(x_i^1, y_i^1; \alpha\right)}{\hat{D}_G^0\left(x_i^0, y_i^0; \alpha\right)} \frac{\hat{D}_G^1\left(x_i^1, y_i^1; \alpha\right)}{\hat{D}_G^1\left(x_i^0, y_i^0; \alpha\right)}\right]^{1/2}, \tag{111}$$

which can be decomposed into the geometric mean of different terms as in (109), with these terms depending on the directional factor α.

If we depart from the additive directional distance function characterizing the production technology (4), $D_T(x, y; -g_x, g_y)$, productivity analysis leads to the Luenberger indicator, which following Balk (1998; p.174) is defined as—considering $g = \left(-g_x, g_y\right)$ to alleviate notation:

$$\begin{aligned} L_T^{0,1}(x_i^1, y_i^1, x_i^0, y_i^0; g) &= D_T^0(x_i^0, y_i^0; g) - D_T^1(x_i^1, y_i^1; g) + \\ &+ \frac{1}{2}\left[\left(D_T^1(x_i^0, y_i^0; g) - D_T^0(x_i^0, y_i^0; g)\right) + \left(D_T^1(x_i^1, y_i^1; g) - D_T^0(x_i^1, y_i^1; g)\right)\right] \\ &= \Delta TE_T(x_i^1, y_i^1, x_i^0, y_i^0; g) + \frac{1}{2}\left[TC_T^{1,0}(x_i^0, y_i^0; g) + TC_T^{1,0}(x_i^1, y_i^1; g)\right]. \end{aligned} \tag{112}$$

Therefore, as in the MPI case, productivity change can be decomposed into technical efficiency change and the average of technological change evaluated at (x_i^0, y_i^0) and (x_i^1, y_i^1). Clearly, if the values of these terms are positive (negative) there has been an increase (decrease) in technical efficiency and technological progress (regress), respectively.[86] However, while the

[86] As it is defined, the larger the numerical value of the Luenberger indicator, the greater the productivity change, extending to the technical efficiency change and technological change. This in contrast to the contemporary definition of profit, technical and allocative inefficiency in (17) where the larger the values, the worse the firm's performance. This justifies adopting ΔTE_T as the notation for change in technical inefficiency *TI*.

values of MPI correspond to the change in average productivities between the base and comparison period−as easily shown in the single input-single output case, an intuitive interpretation of the numerical values of the Luenberger indicator is not readily available. [87, 88]

8.2. Parametric decomposition of total factor productivity

Naturally, as distance functions can be estimated parametrically, they also constitute the building blocks for the measurement of productivity change and its decomposition into basic sources of efficiency change and technical change. This decomposition can be helpful to guide policy if estimated with precision.

Total factor productivity growth is often defined as the rate of growth of output minus the rate of growth in the input usage. Assuming that firm's technology can be represented by a Translog output distance function, Orea (2002) use the following Generalized Malmquist Productivity Index between the base ($t = 0$) and comparison ($t = 1$) periods:

$$\ln G_O = \frac{1}{2}\sum_{m=1}^{M}\left[e_m^1 + e_m^0\right]\cdot\ln\left(\frac{y_m^1}{y_m^0}\right) + \frac{1}{2}\sum_{n=1}^{N}\left[\frac{e_n^1}{\Sigma_{n=1}^{N} e_n^1} + \frac{e_n^0}{\Sigma_{n=1}^{N} e_n^0}\right]\cdot\ln\left(\frac{x_n^1}{x_n^0}\right), \tag{113}$$

where $D(t)$ is short for $D\left(x_i^t, y_i^t, t, \beta\right)$, $e_m^t = \partial \ln D\left(t\right) / \partial \ln y_m$ is the elasticity of the distance function with respect to m-th output, $e_n^t = \partial \ln D\left(t\right) / \partial \ln x_n$ is the elasticity of the distance function with respect to n-th input. This productivity index can be broadly defined as the difference between the weighted average rates of growth of outputs and inputs, where the weights are output and (normalized) input distance elasticities respectively.[89]

The starting point for decomposing this productivity measure is the estimated distance function. Note that the translog specification of the output distance function in (29) can be regarded as a quadratic function in logs. Hence, it is possible to apply Diewert's (1976) Quadratic Identity Lemma. Using this identity, Orea (2002) obtain the following parametric decomposition of this productivity index: [90]

$$\begin{aligned}\ln G_O &= \ln D(1) - \ln D(0) \\ &- \frac{1}{2}\left[\frac{\partial \ln D(1)}{\partial t} + \frac{\partial \ln D(0)}{\partial t}\right] \\ &+ \frac{1}{2}\sum_{n=1}^{N}\left[EE^1\frac{e_n^1}{\Sigma_{n=1}^{N} e_n^1} + EE^0\frac{e_n^0}{\Sigma_{n=1}^{N} e_n^0}\right]\cdot\ln\left(\frac{x_n^1}{x_n^0}\right),\end{aligned} \tag{114}$$

[87] Boussemart et al. (2003) and Balk et al. (2008) study the relationship between the Malmquist productivity indices and Luenberger productivity indicators.

[88] A hybrid definition known as the Malmquist-Luenberger productivity index, defined as the ratio of directional distance functions accounting for undesirable production in environmental studies is discussed below.

[89] Equation (113) is a total factor productivity index because it satisfies the proportionality property (as its weights sum to one), as well as the identity, separability, and monotonicity properties. Notice that the output distance function is homogeneous of degree +1 in the output quantities and, as a result, re-scaling the output elasticities is not necessary.

[90] A similar decomposition was simultaneously introduced by Brümmer et al. (2002) in a continuous time framework.

where $EE^t = -\Sigma_{n=1}^{N} e_n^t - 1$, t = 0, 1, is a measure of firms' economies of scale. Equation (114) provides a meaningful decomposition of a generalized MPI into changes in technical efficiency, technical change and a scale effect.[91] As the output distance function can be viewed as Farrell's output-oriented measure of technical efficiency, $\ln D(1) - \ln D(0)$ 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. Unlike Balk (2001) and (109) above, the contribution of scale economies to productivity change is evaluated without recourse to scale efficiency measures, which are neither bounded for globally increasing, decreasing, or constant returns to scale technologies nor for ray-homogeneous technologies.

It should be pointed out that the above decomposition does not individualize any output or input mix effect as presented in (109). 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 (114) can be in turn decomposed in a *pure* scale effect and a term measuring relative changes in the input mix:

$$\begin{aligned} SE = &\left\{ \frac{1}{2} \sum_{n=1}^{N} \left[EE^1 \frac{e_n^1}{\Sigma_{n=1}^{N} e_n^1} + EE^0 \frac{e_n^0}{\Sigma_{n=1}^{N} e_n^0} \right] \cdot \right\} \ln \left(\frac{\overline{x}^1}{\overline{x}^0} \right) \\ &+ \frac{1}{2} \sum_{n=1}^{N} \left[EE^1 \frac{e_n^1}{\Sigma_{n=1}^{N} e_n^1} + EE^0 \frac{e_n^0}{\Sigma_{n=1}^{N} e_n^0} \right] \cdot \ln \left(\frac{\tilde{x}_n^1}{\tilde{x}_n^0} \right), \end{aligned} \tag{115}$$

where $\overline{x}^t = \Pi_{n=1}^{N} \left(x_n^t \right)^{1/N}$ and $\tilde{x}_n^t = x_n^t / \overline{x}^t$, $t = 0, 1$. A similar output mix effect can be obtained if we decompose the output growth in equation (113) taking into account the *average* change in outputs.

8.3. Flexible functional forms and superlative and exact quantity and price indices

Productivity measurement has been traditionally undertaken through classic value indices using ratio formulations such as the Fisher and Törnqvist definitions, or using differences as the Bennet and Montgomery indicators—Balk (2008). These indices use prices rather than distance functions as aggregators, and can be calculated in a simpler way than their quantity only counterparts when price data are reliable. However, relying on the economic theory approach to index numbers it is possible to decompose them into quantity and price indices through duality. For this purpose, we recall the discussion on Section 6.2 on the choice of functional form, as the alternative parametric specifications represent the link that allows the approximation and decomposition of these classical value indices.

The Fisher and Törnqvist indices are 'superlative' in Diewert's (1976) terminology because they are 'exact' for a flexible aggregator that corresponds, precisely, to specific functional forms such as those previously discussed in Section 4.3. In turn, "exactness" implies that a particular

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

index number can be directly derived from that specific flexible aggregator. In production analysis, a quantity (price) index is superlative if it is exact for a flexible production (cost) function. Here we show the relationship between the production function and the quantity indices, but equivalent relationships hold for the (unit) cost function and price indices.

Consider the following flexible quadratic mean of order *r* production function, nesting many widely used functional forms:

$$f_r(x)=\left(\sum_{n=1}^{N}\sum_{n'=1}^{N}\alpha_{nn'}x_n^{r/2}x_{n'}^{r/2}\right)^{1/r}, \quad r\neq 0, \quad \alpha_{nn'}=\alpha_{n'n}., \tag{116}$$

Balk (2008; p. 50-52) and Hill (2006) show that it is the counterpart specification for the Fisher ideal quantity index for $r = 2$. This index defines as:

$$Q_x^F=\left(\frac{\sum_{n=1}^{N}w_n^0x_n^1}{\sum_{n=1}^{N}w_n^0x_n^0}\frac{\sum_{n=1}^{N}w_n^1x_n^1}{\sum_{n=1}^{N}w_n^1x_n^0}\right)^{1/2}=\left(Q_x^LQ_x^P\right)^{1/2}, \tag{117}$$

where, once again, w_n^t ad x_n^t represent inputs' prices and quantities, and superscripts refer to the reference period, base ($t = 0$), and comparison ($t = 1$)−defining each the Laspeyres Q^L and Paasche Q^P indices. Equivalently, the quadratic unit cost function with prices as arguments $C_r(p)$, $r = 2$, corresponds to the Fisher ideal price index $P_x^F=\left(P_x^LP_x^P\right)^{1/2}$. Alternatively, if *r* tends to zero, (116) takes the form of the translog production function specification, for which the Törnqvist index is exact:

$$Q_x^T=\prod_{n=1}^{N}\left(\frac{x_n^1}{x_n^0}\right)^{\frac{(s_n^0+s_n^1)}{2}}, \tag{118}$$

where $s_n^t=w_n^tx_n^t\Big/\sum_{n=1}^{N}w_n^tx_n^t$, $t = 0, 1$, are the cost shares of the inputs in each period. Again, the unit cost function $C_r(p)$ with $r = 0$ corresponds to the Törnqvist input price index. Finally, if $r = 1$, (116) adopts the generalized-Leontief specification, for which the Walsh quantity index is exact, and a similar case is true for the cost function and Walsh's price index.

Defining now output quantity indices symmetric to (117) and (118), which we denote by Q_y^F and Q_y^T , it is possible to formulate the Fisher and Törnqvist productivity indices:

$$Q^F=Q_y^F/Q_x^F=\left(\frac{\sum_{m=1}^{M}p_m^0y_m^1}{\sum_{m=1}^{N}p_m^0y_m^0}\frac{\sum_{m=1}^{M}p_m^1y_m^1}{\sum_{m=1}^{M}p_m^1y_m^0}\right)^{1/2}\Bigg/\left(\frac{\sum_{n=1}^{N}w_n^0x_n^1}{\sum_{n=1}^{N}w_n^0x_n^0}\frac{\sum_{n=1}^{N}w_n^1x_n^1}{\sum_{n=1}^{N}w_n^1x_n^0}\right)^{1/2}=\left(Q_y^LQ_y^P\right)^{1/2}/\left(Q_x^LQ_x^P\right)^{1/2}, \tag{119}$$

$$Q^T = Q_y^T / Q_x^T = \prod_{m=1}^{M} \left(\frac{y_n^1}{y_n^0} \right)^{\frac{(s_m^0 + s_m^1)}{2}} / \prod_{n=1}^{N} \left(\frac{x_n^1}{x_n^0} \right)^{\frac{(s_n^0 + s_n^1)}{2}}, \tag{120}$$

where p_m^t and y_m^t represent outputs' prices and quantities, and $s_m^t = p_m^t y_m^t \Big/ \sum_{m=1}^{M} p_n^t y_n^t$, $t = 0, 1$, are the revenue shares.

Given these options, this leads to the question about the choice of functional form to characterize the technology though the primal and dual (cost) approaches when undertaking productivity studies in a multiplicative (ratio) setting−for the approach based on differences like the Bennet or Montgomery indicators, see Balk et al. (2004) and Diewert (2005).

Fisher is the preferred formula from an axiomatic perspective because it satisfies a large number of tests (hence the "ideal" denomination), it is bounded by the Laspeyres and Paasche indices, and it has a greater intuitive appeal as it corresponds to their geometric mean. On the other, Törnqvist is widely used in applied economic research due to its underlying translog specification. Finally, Walsh is the only fixed-basket superlative index. Overall, it appears that there is no strong reason to prefer one form from another, as we expect that the empirical results from the two functional forms are similar. This should come as no surprise because superlative index number approximate each other to the second order. However, Hill (2006) shows that this mathematical property does not necessarily result in numerical similarity, as the spread between the largest and smallest superlative indices sometimes (but rarely) exceeds that between the Laspeyres and Paasche indices.

This latter finding may have significant implications in applied research with a policy orientation (i.e., productivity, consumer price indices,…), as there is no single answer as to which superlative index, grounded on economic theory, should be used. Beyond the fact that all of them are easy to compute and approximate each other, this suggests that a combination between the economic theory and axiomatic approaches is needed, favouring the Fisher proposal. But we note that the Laspeyres or Paasche formulations are based on an underlying fixed-coefficient technology, which is a strongly simplifying assumption excluding inputs and outputs substitutability, and imposing constant marginal products (OECD, 2001). On these grounds, the question about the preferred functional form remains open.

8.4. Productivity indices and indicators

To the extent that distance functions represent flexible aggregators accommodating multiple-output multiple-input representations of the production technology, it seems natural to relate the well-known index numbers presented above, (119) and (120) with those that define productivity measures in terms of the generalized (5) and directional distance functions (4) such as the Malmquist productivity index (108) and Luenberger productivity indicator (112) already presented.

Diewert (1992) introduces a specific flexible functional form and, assuming constant returns to scale and that in each period firms competitively maximize revenue given outputs and minimize cost given inputs, shows that the Fisher productivity index (119) is exact to the Malmquist index (108). Alternatively, Caves el al. (1982) show that under equal technological

conditions and optimizing economic behaviour, the same relationship can be obtained through duality theory in terms of the translog distance function and the Törnqvist productivity index (120) .

An exact relationship between the Luenberger productivity indicator (112) based on the directional distance function and a value productivity indicator can be also established—Balk (1998). The indicator corresponds to the Bennet formulation:

$$Q^{B} = \frac{1}{2}\left(\frac{p^{0}}{p^{0}g_{y} + w^{0}g_{x}} + \frac{p^{1}}{p^{1}g_{y} + w^{1}g_{x}} \right)\left(y^{1} - y^{0} \right) - \frac{1}{2}\left(\frac{w^{0}}{p^{0}g_{y} + w^{0}g_{x}} + \frac{w^{1}}{p^{1}g_{y} + w^{1}g_{x}} \right)\left(x^{1} - x^{0} \right), \quad (121)$$

which is a price ratio weighted arithmetic mean of the changes in output and input quantities, using as a normalizing condition that already introduced for the single period profit efficiency (17), by which input and output prices are multiplied by the corresponding elements of the directional vector. In this case, using the quadratic specification, Balk (1998; p. 175) and Chambers (2002) show that the Bennet indicator (121) is exact to the Luenberger indicator (112).

Consequently, formal relationships exist between the Fisher and Törnqvist productivity indices, and the quadratic and translog distance functions underlying the Malmquist productivity index, and between the Bennet indicator and the Luenberger indicator based also on the quadratic distance function. Because of this, it is possible to equivalently quantify productivity change either using the former definitions that require information on quantities and prices−but do not need to estimate underlying technology through the corresponding 'exact' functional forms, or alternatively by using the latter that require only quantitative information−but needs to approximate unobserved technology. One advantage of the latter approach is that by approximating the technology through distance functions it is possible to determine the contribution that efficiency and technological change make to productivity change, as presented in the decompositions of the Malmquist productivity index (108) and the Luenberger indicator (112).

8.5. The decomposition of price-based productivity and economic efficiency change

In all the cases mentioned above, in order to establish the exactness between the classical formulations depending on quantities and prices and the recent definitions requiring only quantities, duality is called upon to substitute observed prices in these superlative indices for the derivative of the distance functions with respect to quantities; i.e., marginal productivities. Therefore, as anticipated already, both paths can be followed when calculating productivity change.

This raises the question about the differences encountered between both sets of definitions, e.g., between the Fisher and Malmquist indices, and between the Bennet and Luenberger indicators. The general explanation for the differences is the divergence between the underlying assumptions made to attain the exactness results and reality. Clearly, the specific flexible

functional form is appropriate in reflecting unobserved technology, the nature of variable returns to scale, and whether firms are capable of realizing an optimal economic behaviour in reality.

With respect to this last comment, firms may fall short from attaining their maximum profitability or profit because of technical and allocative considerations. But while the Malmquist productivity index and Luenberger indicator are able to answer questions about the contribution that efficiency and technological change make to productivity change, they cannot capture deviations due to allocative inefficiency, resulting from the inability of firms to demand and supply optimal quantities of inputs and outputs, respectively. Therefore, extending the overall economic decomposition framework presented in Section 2.3 to the multiple period setting, recent research has explored the possibility of defining the relationship between the superlative productivity indices and indicators and allocative (in)efficiency; i.e., a decomposition of profitability and profit change based on technical and allocative criteria.

For example, departing from the duality relationship (13) and the concept of overall (Nerlovian) profit efficiency (17), Juo et al. (2015) propose a result that, as shown by Balk (2017), decomposes a Bennet indicator into the Luenberger indicator (112) and a price (allocative) term, along with a series of additional components:

$$\begin{aligned} Q^B &= \frac{1}{2}\left(\tilde{p}^0+\tilde{p}^1\right)\left(y^1-y^0\right)-\frac{1}{2}\left(\tilde{w}^0+\tilde{w}^1\right)\left(x^1-x^0\right)= \\ &= L_T^{0,1}(x_i^1,y_i^1,x_i^0,y_i^0;g)+\left(AE_T^0(\tilde{w}^0,\tilde{p}^0,x^0,y^0;g)-AE_T^1(\tilde{w}^1,\tilde{p}^1,x^1,y^1;g)\right)+ \\ &+\left(\tilde{\pi}^1(\tilde{w}^0,\tilde{p}^0)-\tilde{\pi}^0(\tilde{w}^1,\tilde{p}^1)\right)-\left(\frac{1}{2}\left(\tilde{p}^1-\tilde{p}^0\right)\left(y^0+y^1\right)-\frac{1}{2}\left(\tilde{w}^1-\tilde{w}^0\right)\left(x^0+x^1\right)\right)- \\ &-\left(\frac{1}{2}\left[\left(D_T^1(x_i^0,y_i^0;g)-D_T^0(x_i^0,y_i^0;g)\right)+\left(D_T^1(x_i^1,y_i^1;g)-D_T^0(x_i^1,y_i^1;g)\right)\right]\right) \\ &= L(x_i^1,y_i^1,x_i^0,y_i^0;g)+\Delta AE^{0,1}+\Delta\tilde{\pi}^{0,1}-P^B-TC^{0.1}(x_i^0,y_i^0,x_i^1,y_i^1;g), \end{aligned} \tag{122}$$

where $\tilde{p}^t$ and $\tilde{w}^t$, t = 0, 1 are normalized prices as in (121), $\Delta AE^{0,1}$ is the additive version of allocative efficiency change, $\Delta\tilde{\pi}^{0,1}$ is the change in normalized profit, P^B is the Bennet price (recovery) indicator, and $TC^{0.1}(x_i^0,y_i^0,x_i^1,y_i^1;g)$ reflects technical change as in (112). The last three terms correspond to a residual that Juo et al. (2015) call 'price effect' (*PE*), as technical change is subtracted.

Symmetrically, based on the duality relationship (14) and the overall (RD) profitability efficiency (18), Zofío and Prieto (2006) propose the following decomposition of the Fisher index (119) into the Malmquist index and allocative terms:

$$\begin{aligned} Q^F &= \left[Q^L Q^P\right]^{1/2} = \left[\frac{p^0y_i^1/w^0x_i^1}{p^0y_i^0/w^0x_i^0}\cdot\frac{p^1y_i^1/w^1x_i^1}{p^1y_i^0/w^1x_i^0}\right]^{1/2}= \\ &= \left[\frac{\hat{D}_G^0(x_i^1,y_i^1;\alpha)}{\hat{D}_G^0(x_i^0,y_i^0;\alpha)}\cdot\frac{\hat{D}_G^1(x_i^1,y_i^1;\alpha)}{\hat{D}_G^1(x_i^0,y_i^0;\alpha)}\right]^{1/2}\cdot\left[\frac{AE_G^1(p^1,w^1,x_i^1,y_i^1;\alpha)}{AE_G^0(p^0,w^0,x_i^0,y_i^0;\alpha)}\cdot\frac{A_G^0(p^0,w^0,x_i^1,y_i^1;\alpha)}{A_G^1(p^1,w^1,x_i^0,y_i^0;\alpha)}\right]^{1/2}= \\ &= \hat{M}_G^{0,1}(x_i^0,y_i^0,x_i^1,y_i^1;\alpha)\cdot AE_G^{0.1}(p^0,w^0,p^1,w^1,x_i^0,y_i^0,x_i^1,y_i^1;\alpha)\cdot\Delta A^{0,1}(p^0,w^0,p^1,w^1,x_i^0,y_i^0,x_i^1,y_i^1;\alpha), \end{aligned} \tag{123}$$

where $AE_G^{0,1}$ represents allocative efficiency change, $\Delta A^{0,1}$ is again a price residual term that corresponds to the ratio of mixed period allocative performances. The Malmquist index itself can be decomposed as in (109) in order to reveal the technological sources of productivity change.

In both cases, as acknowledged by Balk (2017) for the Bennet indicator suggested by Juo et al. (2015), and Zofío and Prieto (2006) themselves for the Fisher index, some of the terms involve deviations from optimal economic behaviour in cross periods, rendering their interpretation difficult. But again, when the underlying assumptions do not hold; i.e., the counterpart flexible functional forms characterizing the technology do not correspond to the Translog or Quadratic specifications, do not have time-invariant second order coefficients, and the firm is not allocative efficient in the g and α directions, the difference between the Fisher and Malmquist indices and the Bennet and Luenberger indicators are caused by unmet assumptions. Further research in this field is necessary to improve the interpretation of the existing and newly proposed decompositions.

Finally, we conclude this section establishing that, following Balk (1998, p.175-178), and based on the duality relationship (13) and the above assumptions, the change in profit can be consistently decomposed into the dual Luenberger quantity indicator and the Bennet price (recovery) indicator:

$$\begin{aligned}(\tilde{p}^1 y^1 - \tilde{w}^1 x_i^1) - (\tilde{p}^0 y^0 - \tilde{w}^0 x_i^0) = L_T^{0,1}(x_i^1, y_i^1, x_i^0, y_i^0; g) + \\ + \frac{1}{2}\left(\tilde{p}^1 - \tilde{p}^0\right)\left(y^0 + y^1\right) - \frac{1}{2}\left(\tilde{w}^1 - \tilde{w}^0\right)\left(x^0 + x^1\right),\end{aligned} \tag{124}$$

and a similar relationship could be in principle established for profitability change based on the Malmquist productivity index and the corresponding price index counterpart.

8.6. Environmental productivity: The Malmquist and Malmquist-Luenberger indices

Extending Section 7.2 on environmental efficiency measurement to a time series setting, it is possible to define environmental productivity change relying on the multiplicative Malmquist productivity index and the additive Luenberger indicator. This simply requires substituting the environmental distance functions accounting for the undesirable outputs (89) and (90) for the standard ones (4) and (5).

Following the structure of section 8.1 when panel data is available it is straightforward to calculate a firm's environmental productivity change from the base period (x_i^0, y_i^0, b_i^0) to the comparison period (x_i^1, y_i^1, b_i^1) depending on the choice of distance function. The multiplicative environmental Malmquist productivity index, EMPI, accounting for undesirable outputs corresponds to the following definition:

$$\hat{\mathrm{M}}_G^s(x_i^1, y_i^1, b_i^1, x_i^0, y_i^0, b_i^0, ;\alpha) = \frac{\hat{D}_G^s\left(x_i^1, y_i^1, b_i^1; \alpha\right)}{\hat{D}_G^s\left(x_i^0, y_i^0, b_i^0; \alpha\right)}. \tag{125}$$

Here, once again, the EMPI is the ratio of two generalized distance functions defined under either the base (s=0) or comparison period (s=1) virtual (cone) technologies characterized by

constant returns to scale, CRS. Mirroring (109) it is possible to decompose the MPI into a series of indices associated to the different terms contributing to productivity change, either by adopting an output ($\alpha = 1$), input ($\alpha = 0$), or a generalized orientation ($0 < \alpha < 1$). As before, values greater (smaller) than one identify productivity growth (decrease), technical efficiency gains (losses) and technological progress (regress). On this occasion, the geometric mean corresponds to:

$$\hat{\mathrm{M}}_G^{0,1}(x_i^0, \mathrm{y}_i^0, b_i^0, x_i^1, y_i^1, b_i^1; \alpha) = \left[\frac{\hat{D}_G^0\left(x_i^1, y_i^1, b_i^1; \alpha\right)}{\hat{D}_G^0\left(x_i^0, y_i^0, b_i^0; \alpha\right)} \cdot \frac{\hat{D}_G^1\left(x_i^1, y_i^1, b_i^1; \alpha\right)}{\hat{D}_G^1\left(x_i^0, y_i^0, b_i^0; \alpha\right)} \right]^{1/2} \tag{126}$$

Alternatively, it is also straightforward to rely on the environmental directional distance function and define the corresponding environmental Luenberger indicator—considering once again $g = \left(g_y, -g_b\right)$ to alleviate notation:

$$\begin{aligned} L_T^{0,1}(x_i^1, y_i^1, b^1, x_i^0, y_i^0, b^0; g) &= D_T^0(x_i^0, y_i^0, b^0; g) - D_T^1(x_i^1, y_i^1, b^1; g) + \\ &+ \frac{1}{2}\left[\left(D_T^1(x_i^0, y_i^0, b^0; g) - D_T^0(x_i^0, y_i^0, b^0; g) \right) + \left(D_T^1(x_i^1, y_i^1, b^1; g) - D_T^0(x_i^1, y_i^1, b^1; g) \right) \right] \\ &= \Delta TE_T(x_i^1, y_i^1, b^1, x_i^0, y_i^0, b^0; g) + \frac{1}{2}\left[TC_T^{1,0}(x_i^0, y_i^0, b^0; g) + TC_T^{1,0}(x_i^1, y_i^1, b^1; g) \right], \end{aligned} \tag{127}$$

with the analogous technical efficiency change and technological change interpretations, evaluated at (x_i^0, y_i^0, b_i^0) and (x_i^1, y_i^1, b_i^1). In this case a positive value (negative) signals whether there has been environmental productivity growth (decline), technical efficiency gains (losses) or technological progress (regress)—although caution must be exerted when defining the production technology as discussed in section 7.2 and what follows, so as to ensure a consistent interpretation of the environmental technical efficiency term.

However, the literature on environmental productivity change, rather than considering the above two extensions of the existing Malmquist and Luenberger proposals, has embraced a new definition termed the Malmquist-Luenberger productivity index, which defines the ratio of two directional distance functions, and yields a multiplicative decomposition into two mutually exclusive technical efficiency change and technological change components. The index, introduced by Chung et al. (1997), has the advantage of combining a multiplicative definition that renders it comparable with the popular Malmquist index, and incorporates the flexibility that the directional distance function offers when treating desirable and undesirable outputs asymmetrically. The disadvantage is that, in contrast to its Malmquist counterpart based on the environmental generalized distance function (90), its properties and interpretation remain an issue.

Following Färe et al. (2001), the Malmquist-Luenberger index, MLI, with orientation $g = \left(g_y, -g_b\right) = (y, -b)$, and based on period s technology is:

$$ML^s(x_i^1, y_i^1, b^1, x_i^0, y_i^0, b^0; g) = \frac{1 + D_T^s\left(x^0, y^0, b^0; y^0, -b^0\right)}{1 + D_T^s\left(x^1, y^1, b^1; y^1, -b^1\right)}, \quad s = 0,1. \tag{128}$$

As for its Malmquist and Luenberger counterparts, the MLI index may be decomposed into two terms corresponding to efficiency change and technical change as follows:

$$ML^0(x_i^1, y_i^1, b^1, x_i^0, y_i^0, b^0; g) = \frac{1+D_T^0\left(x^0, y^0, b^0; y^0, -b^0\right)}{1+D_T^1\left(x^1, y^1, b^1; y^1, -b^1\right)} \cdot \frac{1+D_T^1\left(x^1, y^1, b^1; y^1, -b^1\right)}{1+D_T^0\left(x^1, y^1, b^1; y^1, -b^0\right)} = \\ = EC_T(x_i^0, y_i^0, b_i^0, x_i^1, y_i^1, b_i^1) \cdot TC_T^{0,1}(x_i^1, y_i^1, b_i^1), \tag{129}$$

and

$$ML^1(x_i^1, y_i^1, b^1, x_i^0, y_i^0, b^0; g) = \frac{1+D_T^0\left(x^0, y^0, b^0; y^0, -b^0\right)}{1+D_T^1\left(x^1, y^1, b^1; y^1, -b^1\right)} \cdot \frac{1+D_T^1\left(x^0, y^0, b^0; y^0, -b^0\right)}{1+D_T^0\left(x^0, y^0, b^0; y^0, -b^0\right)} = \\ = EC_T(x_i^0, y_i^0, b_i^0, x_i^1, y_i^1, b_i^1) \cdot TC_T^{0.1}(x_i^0, y_i^0, b_i^0). \tag{130}$$

One more time, the geometric mean of the two period indices, corresponding to $ML^{0,1} = \left(ML^0 \cdot ML^1\right)^{1/2}$, avoids the arbitrary choice of a reference technology. The index can be decomposed into the same two components, accounting for efficiency change and technical change:

$$ML^{0,1} = \left(ML^0 \cdot ML^1\right)^{1/2} = \frac{1+D_T^0\left(x^0, y^0, b^0; y^0, -b^0\right)}{1+D_T^1\left(x^1, y^1, b^1; y^1, -b^1\right)} \cdot \left[TC_T^{0,1}(x_i^1, y_i^1, b_i^1) \cdot TC_T^{0,1}(x_i^0, y_i^0, b_i^0)\right]^{1/2} \tag{131}$$

We now recall the initial characterization of the production technology through axioms (A1)−(A7) discussed in Section 7.2. Aparicio et al. (2017) show that operationalizing the DEA model (91) through the directional distance function (89) not only constitutes a compromise that solves the drawbacks associated to the weak disposability of undesirable outputs−and those resulting from considering them as strongly disposable inputs, but also the numerical inconsistency that the technical change term of the ML index exhibits under the original (A1)−(A6) axiomatic framework−e.g., as in Färe et al. (2001), Kumar (2006), among others. This inconsistency is related to the fact that the actual shift in the production possibility set may not be correctly captured by the numerical value of $TC_T^{0.1}$ above. For example, as illustrated in Figure 3a for (x_B, y_B, b_B), where environmentally friendly technical progress takes place as more desirable outputs are produced with fewer undesirable outputs, technical change is nevertheless associated to $TC_T^{0,1}(x_i^1, y_i^1, b_i^1) < 1$ in (129), indicating unreal technological regress—since $D_T^1\left(x^1, y^1, b^1; y^1, -b^1\right) = 0$ and $D_T^0\left(x^1, y^1, b^1; y^1, -b^1\right) > 0$.[92] Moreover, the Malmquist-Luenberger index itself—through the technological change term—may yield wrong numerical values. These authors survey the large number of studies that adopt the original axiomatic frameworks and whose results, managerial advice and policy recommendations are compromised by this weakness.

In a multi-period setting, adopting the extended axiomatic framework (A1)−(A7) provides a solution to the inconsistency problem by ensuring that production technologies are nested−with $\overline{P}^0(x) \subseteq \overline{P}^1(x)$, if the upper bound for the production of undesirable outputs is defined for all

[92] On the other hand, note that the counterpart $TC_T^{0,1}(x_i^0, y_i^0, b_i^0)$ in (130) is infeasible.

periods by $\bar{b}^0(x) = \bar{b}^1(x) = \max_{s=0,1}\{b^s\}$; i.e., the maximum observed quantity of undesirable outputs.

The production possibility set is now represented in Figure 3b, where the environmental technologies are nested. Therefore, the new axiom (A7) ensures both that the efficient observations conforming the production frontier do not yield non-positive shadow prices and, most importantly, that the numerical value of the technological change is consistent with the shift in the environmental production frontier.

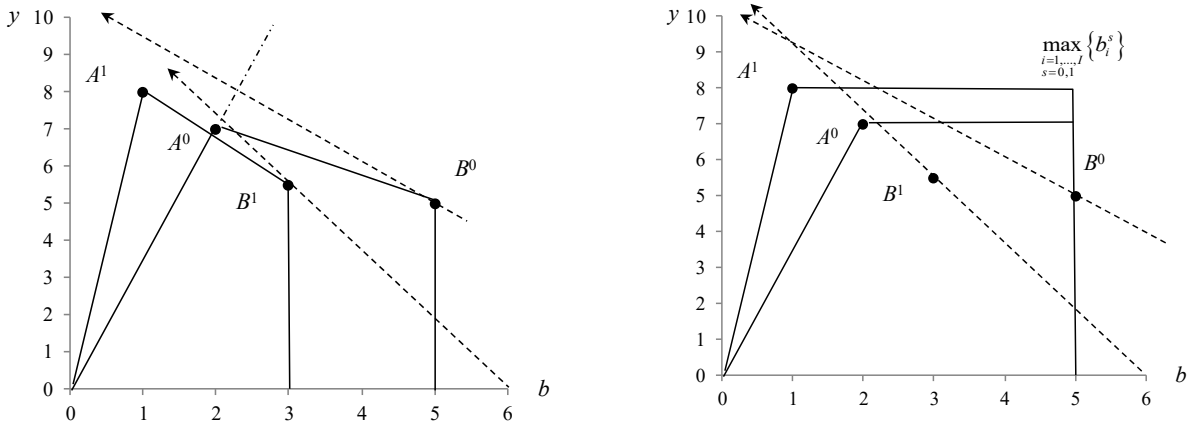


Figure 3a-b. Environmental productions sets under the standard and new axiomatic framework.

Again, the Malmquist, Luenberger, and Malmquist-Luenberger productivity formulations−(125), (127) and (128)−can be computed by relying on non-parametric and parametric techniques. However, while Aparicio et al. (2017) implement the activity analysis DEA modelling, an initial parametric proposal is to be developed.

9. Dynamic efficiency measurement

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. This issue has been little examined in the efficiency measurement literature but has recently become an important concern due to the increasing accessibility to panel data sets. For a more comprehensive review of this literature see Emvalomatis (2009).

Two different approaches have been used in the literature to incorporate the dynamic nature of the decision-making process into efficiency analyses. One approach is to use *reduced-form* models that do not define explicitly a mathematical representation of dynamic behaviour of the firm, but instead incorporate the implications of such an underlying model into static models of efficiency measurement. Some dynamic aspects of firm behaviour are accommodated in these models without imposing strong assumptions on the data. The second approach is the use of *structural models* that make explicit assumptions regarding the objective of the firm and on a rule for forming expectations with respect to future input prices and technological advances. On this occasion, the economic objective of the firm enhances that presented in Section 2, as it corresponds to the maximization of the sum of discounted profit flows−or the minimization of discounted cost−under dynamic constraints.

9.1. Reduced-form models

Stochastic frontier models make available two alternative approaches to deal with time dependent inefficiencies. The most common approach estimates the temporal pattern of the variation in inefficiency by using deterministic specifications of time. See, for instance, Kumbhakar (1991), Battese and Coelli (1992), Lee and Schmidt (1993), and Cornwell et al. (1990). These models suffer from the problem of imposing arbitrary restrictions on the short-run efficiency as well as being unable to model the dynamic nature of the decision-making process.

A more recent approach relates to the dynamic behaviour of inefficiency by considering models that estimate long-run efficiency. Ahn et al. (2000) provide a model of dynamic inefficiency based as far as possible on the fundamental aspects of existing stochastic frontier models. Huang and Chen (2009) formulate a multi-period forward-looking rational expectations model on the evolution of the technical inefficiency level, which accommodates the dynamic efficiency model pioneered by Ahn et al. (2000). These authors recognize a persistence effect of firms' inefficiency over time and specify its evolution as an autoregressive process. A criticism of the model is that it does not restrict $u_i^t=\ln(TE_i^t)$ to be nonnegative. Another criticism is that it does not allow for environmental variables to play a role in the determination of the efficiency levels.

A related model was developed by Tsionas (2006). It departs from a typical stochastic production frontier of the following form:

$$\ln y_i^t = \ln f\left(x_i^t, t, \beta\right) + v_i^t + \ln(TE_i^t), \tag{132}$$

where TE_i^t is the usual technical efficiency of firm i in period t. Contrary to Ahn et al. (2000), this model avoids the complications inherent in the specification of autoregressive processes on non-negative variables because one-to-one mapping from the unit interval to the real line is used to convert the technical efficiency term into an autoregressive form:

$$\begin{aligned} s_{it} &= z'^{t}_{i}\,\delta + \rho s_i^{t-1} + \xi_i^{t1}, \quad \text{for } t = 2..., T, \\ s_{i1} &= z'^{t}_{i}\,\delta/(1-\rho) + \xi_i^{1}, \quad \text{for } t = 1, \end{aligned} \tag{133}$$

where $s_i^t = \ln\left(-\ln TE_i^t\right)$. Alternatively, Emvalomatis (2009) and Emvalomatis et al. (2011) use the inverse of the logistic function for the transformation. More precisely, they define $s_i^t = \ln\left(TE_i^t / \left(1-TE_i^t\right)\right)$ 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.[93] Thus, the distinguishing feature of (133) is that past values of efficiency determine the value of TE_i^t.

Estimating dynamic stochastic frontier models is far from simple. The dynamic model in Ahn et al. (2000) is estimated by GMM, with a very large number of instruments in some models. While Tsionas (2006) estimate the model using Bayesian techniques, Emvalomatis et al. (2011) use Kalman filtering techniques and proceed to estimation by maximum likelihood.

Emvalomatis (2012), Galán and Pollitt (2014) and Galán et al. (2015) studied later on the effect of including unobserved heterogeneity in dynamic inefficiency models. Emvalomatis (2012) pointed out that an estimate of ρ could be inflated due to the presence of unobserved heterogeneity in the sample. If there is heterogeneity in the production function and it is ignored, the model will interpret part of it as inefficiency. The result will be an upward bias of the estimate of ρ, as this parameter will now measure the persistence not only of inefficiency, but also of the firm effects. In order to control for this issue, this author includes a time-invariant firm effect, α_i, in the production function (1). Galán et al. (2015) model unobserved heterogeneity through a heterogeneous persistence parameter, i.e., they estimate a firm-specific ρ_i, and related ρ_i to differences in the adjustment costs among firms. Galán and Pollitt (2014) examine both sources of unobserved heterogeneity, α_i and ρ_i. The findings of all these papers reveal high biases in the efficiency estimations when unobserved factors and unobserved differences in the persistence parameters are not considered.

9.2. Structural dynamic models

A key feature of explicit structural models is that they make explicit assumptions on 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 Emvalomatis, 2009; p. 30): [94]

$$\begin{aligned} J &= \max_{I,x} E^t \left\{ \int_0^\infty e^{-\rho t} g(y,x,k,I) dt \right\} \\ s.t. \quad & \dot{k} = I - \delta k \\ & (x,I) \in L(y;k) \\ & given\ k \end{aligned} \tag{134}$$

In this formulation, the objective of the firm is to maximize the expected discounted flow of an instantaneous reward function g over time (e.g., the profit function or the negative cost function). The set of inputs is divided into two subsets, variable (x) and quasi-fixed (k) inputs. The choice variables are the levels of variable inputs to be employed and the level of investment in

[93] The s_{it} process is stationary provided $|\rho|<1$. Stationarity of the s series implies that the expected value of s in the long run is the same for all firms. Given the one-to-one transformation from s to TE, this steady state value of s is directly translated to a long-run expected value for the technical efficiency scores. Any deviation from this long-run expected value could be attributed to random noise, suboptimal decision making, or to the different stages in the adjustment process of firms towards their long-run equilibrium captured by the data. In other words, the model assumes that there are no systematic differences in the efficiency scores of different firms. In the long run the efficiency scores should have a common distribution around the long-run expected value.

[94] This subsection heavily relies on Emvalomatis' (2009) thesis.

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 an input requirements set. Given the level of quasi-fixed inputs, this set describes the vectors of outputs that can be produced from a given vector of variable inputs and gross investment. Adjustment costs are not observed, but are implicit in the above formulation once investment is included as a technology driver in (134). In a parametric approach, the representation of technology could be given in terms of a distance function $D_T(y,x,k,I)$. 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.

The production (cost) function formulation of the adjustment cost theory of investment was introduced by Lucas (1967), while Treadway (1969, 1970) and Morrison (1992) are early contributors to the adjustment cost literature. The possibility of inefficiency at the firm level is not considered in all these studies. The objective of this line of literature is not only to generate estimates of the parameters of a production (cost) function augmented to allow for adjustment costs, but primarily to provide a framework for studying the response of firms to exogenous conditions, mainly price changes. Econometric estimation of such dynamic models is difficult since the optimal control problem rarely has a closed-form solution and flexible functional forms require the estimation of a very large number of parameters. Another challenge in these dynamic models is the way firms' managers form their expectations about future prices.

The issue of dynamic efficiency measurement has been initially addressed in a non-parametric setting. This is because non-parametric methods do not require explicitly solving the firm's optimization problem. As is the case with the static framework, a statement about the objective of the firm is enough to form the piecewise linear efficient frontier.

9.2.1. Non-parametric dynamic efficiency models[95]

While Sengupta (1995) introduced the first-order conditions of the dynamic optimization problem into the DEA models, Nemoto and Goto (1999, 2003) incorporate the adjustment costs into their model by considering quasi-fixed inputs as outputs in the current period, while treating them as inputs in the next period.[96] Therefore, they construct an input requirements set similar to the one in (134), with k^{t-1} replacing investment.

Silva and Stefanou (2007) proposed hyperbolic measures of dynamic efficiency based on their earlier nonparametric dynamic dual cost approach to production analysis (Silva and Stefanou, 2003). They propose temporal efficiency measures, in the sense that they measure the efficiency of a firm at specific locations on the adjustment path of the firm. This temporal nature of efficiency seems to be the appropriate way to measure efficiency in dynamic models because firms operate with a view to the long-run, although decisions are made in the short-run taking into account the information available at each point in time. Later on, Silva and Oude Lansink (2009) develop a

[95] See Fallah-Fini et al (2014) For a more comprehensive review of the literature on dynamic inefficiency measurement in non-parametric setting.

[96] Also using a network approach, Chen (2009), Chen and Van Dalen (2010) and Skevas et al. (2012) propose a dynamic DEA model where intermediate outputs in the current period may affect future output. In this group of studies, it is also important to mention Färe and Grosskopf (1996), Tone and Tsutsui (2010, 2014), and Sueyoshi and Sekitani (2005).

dynamic specification of the input directional distance function and show that this function fully characterizes the input requirement set in a dynamic setting, being thus an alternative primal representation of the adjustment cost production technology. Silva and Oude Lansink (2013) extend the adjustment cost framework to the directional distance function and its dual cost function.[97] Finally, Kapelko et al. (2014) employ a non-parametric DEA approach to measuring dynamic inefficiency in the adjustment-cost technology framework. However, unlike previous works, they make a full decomposition of dynamic cost inefficiency into technical, scale and allocative inefficiency in the directional distance function context. In particular, they assume that each firm minimizes the flow of future costs over time, subject to an adjustment-cost technology that is modelled using the following DEA specification of a dynamic directional input distance function:

$$
\begin{aligned}
&\vec{D}_i = (y, K, x, I; g_x, g_I) = \max_{\beta,\gamma} \beta \\
&s.t. \\
&y \le \sum_{j=1}^{J} \gamma^j y^j, \\
&\sum_{j=1}^{J} \gamma^j x^j \le x - \beta g_x, \\
&I + \beta g_I - \delta K \le \sum_{j=1}^{J} \gamma^j \left(I^j - \delta K^j \right), \\
&\gamma^j \ge 0,\ j = 1, ..., J,
\end{aligned}
\tag{135}
$$

where β is a measure of dynamic technical inefficiency.[98] It represents the maximum contraction of the variable input vector x in the direction of $-g_x$ and simultaneous maximum expansion of the dynamic factor vector I in the direction of g_I. The dynamic directional input distance function thus measures the distance of x and I to the frontier in the direction defined by the directional vector $(-g_x, g_I)$. As usual, a scale efficiency (SE) measure can be obtained once variable and constant returns to scale specifications of the above distance function are estimated. From their intertemporal cost minimization problem, they define the following dynamic cost inefficiency (OE) measure:

$$
OE = \frac{wx + c \cdot K + W_K(\cdot)(I - \delta K) - rW(y, K, w, c)}{wg_x - W_K(\cdot) g_I},
\tag{136}
$$

where $W_K(\cdot)$ is the shadow value of quasi-fixed inputs, $W(\cdot)$ represents the discounted flow of costs in all future time periods, d is the depreciation rate, w and c are the price of variable and quasi-fixed inputs, and r is the discount rate. This measure is the normalized deviation between the shadow cost of the actual choices and the minimum shadow cost ($rW(y,K,w,c)$). The normalization is the shadow value of the directional vector, making the dynamic cost inefficiency a unit-free measure. The overall dynamic cost inefficiency relates to the firm's ability to minimize production costs in order to produce a given level of output.

[97] Silva and Oude Lansink (2009, 2013) has been recently published as Silva et al (2015).

[98] This model has been recently extended to multi-directional inefficiency analysis in the paper by Kapelko and Oude Lansink (2017).

Finally, dynamic overall cost inefficiency is decomposed into the contributions of technical inefficiency under variable returns to scale, scale inefficiency (*SE*) and a residual term defined as allocative inefficiency (*AE*) that refers to the firm's ability to choose the optimal mix of variable and dynamic factors, given their respective prices, i.e., the mix which minimizes long-run costs:

$$OE_i = D_T(y, K, x, I; -g_x, g_I) + SE_i + AE_i \tag{137}$$

9.2.2. Parametric dynamic efficiency models

Only recently we have observed a number of important parametric contributions to dynamic inefficiency modelling in the above adjustment-cost technology framework. Rungsuriyawiboon and Stefanou (2007) conduct the first study where allocative and technical efficiency measures were generated in an intertemporal decision-making framework. They develop a dynamic shadow-cost approach that does not specify or estimate the production technology directly.

Serra et al. (2011) used the adjustment cost framework of Silva and Lansink (2013) based on production technology to estimate dynamic efficiency, but instead of DEA they carried out a parametric estimation generalizing the static input-oriented directional distance function introduced by Färe et al. (2005). Tovar and Wall (2014) have recently use parametric techniques to estimate an input-oriented directional distance function (and a stochastic cost frontier) to measure dynamic technical efficiency for a set of Spanish port authorities. Stochastic estimation is accomplished by maximum likelihood procedures.

The input-oriented dynamic directional distance function used in both papers can be written as

$$\vec{D} = (y, K, x, I; -g_x, g_I) = \max\left\{\beta : (x - \beta g_x, I + \beta g_I) \in V(y : K)\right\}, \tag{138}$$

where *V(y,K)* is the production input requirement set defined as $V(y,K)=\{(xI)$: (x,I) can produce y given $K\}$. Again, this directional distance function represents the maximum contraction of variable inputs and the maximum expansion of investments that keeps the combination of variable inputs and investments inside the input requirement set, while maintaining outputs and quasi-fixed input constant. The choice of directional vector is normally made to simplify the interpretation of the technical inefficiency scores. Again, with the direction $(-g_x, g_I) = (-1, 1)$, the resulting inefficiency score has the nice interpretation in that it represents the equiproportional reduction in variable inputs and simultaneous expansion of investments that can be made. In this context, the choice of directional vector is, generally speaking, a matter of convenience.

Following Färe et al. (2005) and Serra et al. (2011), Tovar and Wall (2016) use a quadratic functional form for the directional distance function. This has the advantage that it is easy to impose parametric restrictions so that it satisfies the translation property.

$$\vec{D}(y, K, x - \lambda g_x, I + \lambda g_I; -g_x, g_I) = \vec{D}(y, K, x, I; -g_x, g_I) - \lambda. \tag{139}$$

This simply states that if investment is expanded by αg_I and input contracted by $-\alpha g_x$, the value of the distance function will be reduced by λ. Setting $(g_x, g_I) = (1, 1)$−dropping the negative sign for notational simplicity, the dynamic directional distance can be expressed as:

$$\vec{D}(y,K,x,I;-1,1) = a_0 + a_y y + a_K K + a_x x + a_I I + \tfrac{1}{2}a_{yy}y^2 + \tfrac{1}{2}a_{KK}K^2 + \tfrac{1}{2}a_{xx}x^2 + \tfrac{1}{2}a_{II}I^2 + a_{yK}yK + a_{yx}yx + a_{yI}yI + a_{Kx}Kx + a_{KI}KI + a_{xI}xI. \quad (140)$$

And the parameter restrictions associated to the translation property are simple to impose: $a_I - a_x + 1 = a_{Kx} - a_{Kx} = a_{yI} - a_{yx} = a_{xI} - a_{xx} = a_{II} - a_{Ix} = 0$.

To estimate the directional distance function using stochastic frontier techniques, the stochastic specification for firm *i* takes the form:

$$0 = \vec{D}(y_i, K_i, x_i, I_i; 1, 1) + v_i - u_i, \quad (141)$$

where $v_i \sim N(0, \sigma_v)$ is white noise and $u_i \sim N(0, \sigma_u)$ is a one-sided term measuring firms' technical inefficiency. The translation property allows us to replace the constant dependent variable with a firm-specific variable. Indeed, equation (141) can be rewritten using the translation property as:

$$\lambda_i = \vec{D}(y_i, K_i, x_i - \lambda_i, I_i + \lambda_i; 1, 1) + v_i - u_i, \quad (142)$$

where the function $\vec{D}(y_i, K_i, x_i - \lambda_i, I_i + \lambda_i; 1, 1)$ is the quadratic form in **Error! Reference source not found.** with λ_i subtracted from variable inputs and added to investment. The above three papers set $\lambda_i = I_i$. Estimating the above directional distance function only provides estimates of technical inefficiency. To get cost efficiency scores in a dynamic framework, Serra et al. (2011) and Tovar and Wall (2016) propose estimating the following (quadratic) cost frontier model:

$$C_i = rW(y,K,w,c) - W_K(\cdot)\dot{K} + \varepsilon_i + \gamma_i, \quad (143)$$

where C_i is observed cost (normalized by a variable input price), $W(\cdot)$ is optimum cost, $W_k(\cdot)$ is its derivative with respect to the capital stock; $\varepsilon_i \sim N(0, \sigma_\varepsilon)$ is white noise and $\gamma_i \sim N(0, \sigma_\gamma)$ is a one-sided term measuring firms' cost inefficiency.

In a similar fashion to the non-parametric setting, the dynamic directional distance function **Error! Reference source not found.** allows estimating technical inefficiency of both variable and quasi-fixed inputs. The parametric dynamic cost model (143) allows estimating the dynamic cost inefficiency defined in (136) as the difference between the observed shadow cost of input use and the minimum shadow cost, normalized by the shadow value of the direction vector. Finally, an allocative inefficiency score can be obtained as the difference between dynamic cost inefficiency and dynamic technical inefficiency.

To conclude this section, and regarding the productivity decompositions in Section 8, it is worth mentioning that there is also a literature developing dynamic productivity growth measures using both parametric and nonparametric techniques. See for instance Oude Lansink et al. (2015), Kapelko et al. (2015), Kapelko (2017), and Kapelko et al. (2017).

10. Concluding remarks

This paper serves as guide to economic efficiency and productivity evaluation from an economic perspective and intends to make the reader aware of the different alternatives available for choice when undertaking research in the field. The analytical framework relies on the most general models and up to date representations of the production technology and economic performance through directional and generalized distance functions, nesting the traditional approaches well-known in the literature, while complementing them with current issues related to their empirical implementation. This allows us to introduce the following decision structure, corresponding to a flow of relevant issues that can be summarized in the following Q&A:

- What is the economic objective (rationality) of firms given the market structure and regulatory constraints? To minimize cost or maximize revenue, profit or profitability, avoiding technical and allocative inefficiencies.

- What is the corresponding (primal) analytical framework based on duality theory? Input, output, directional or generalized distance functions, representing technical efficiency, and allowing the decomposition of overall economic efficiency in static and dynamic settings.

- What is the most appropriate method for the empirical analysis? Either non-parametric (DEA), semi-parametric, or parametric (SFA) techniques, depending on the likelihood of noise and measurement errors, suitability of the method to represent firms' technology, number of decision variables and observations (degrees of freedom), etcetera.

- What are critical issues when deciding upon the final model? Assessing dimensionality requirements for reliable results, using appropriate statistical methods which identify outliers, selecting the most relevant variables to be included in the final model, and choosing the most adequate functional form best suited to the technological and economic objective of the firm.

- What are the contextual (non-discretionary) variables conditioning firms' technological and economic performance needed to be taken into account? Consideration of one or two stage methods including these variables which is dependent upon whether they are relevant when defining the reference frontier (internal factors) or mainly affect individual efficiency (external factors). Additionally, grant consideration to the statistical properties of the efficiency estimator in the alternative approaches.

- What are the relevant estimation issues in mathematical programming and regression analyses? Consideration of endogeneity issues and the methods available to address them, as well as the specific choice of directional vectors. This is particularly so in parametric settings through the imposition of the necessary homogeneity properties in flexible functional forms.

- What are the recent and significant extensions of the basic model? Those requiring a large qualification of the technology to account for risk and uncertainty, plus undesirable or bad outputs in environmental studies.

- What are the panel data and dynamic extensions of the basic model? Total factor productivity analysis through Malmquist indices and Luenberger indicators. Dynamic efficiency analysis in the short and long term incorporating rigidities, fixed and variable inputs, persisting inefficiency, etcetera.

In this scheme we stress the importance of choosing a suitable analytical framework that is in accordance with the industry characteristics and the restrictions that firms face, 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, in an unconstrained setting, is assumed to maximize profit or profitability, both of which can be related to cost minimization and revenue maximization. Based on duality, the choice of economic objective is followed by the appropriate counterpart characterization of the production technology, enabling the decomposition of economic efficiency in a consistent way. Given that distance functions can be interpreted as measures of technical efficiency, the difference between observed cost, revenue, profit or profitability, and the optimally observed benchmark, can be decomposed according to technical and allocative criteria. Distance functions themselves represent the main variables needed to benchmark firms, industries and economies. Multilateral comparisons are possible both across firms and time, defining indices and indicators based on them, such as the Malmquist and Luenberger formulations. Duality theory is also the cornerstone when decomposing traditional indices of profit and profitability change in quantity and price indices and indicators.

It is important to remark that the underlying theoretical economic framework presented in Section 2 is common to both approaches and, therefore, mathematical programming and regression techniques are equally suited to undertake empirical analyses. Particularly, and related to the reference economic framework, we present the main methods for decomposing economic efficiency into technical and allocative terms. Additionally, we comment on recent contributions that stress the importance of homotheticity and separability in correctly interpreting the standard (Farrell) radial distance functions as measures of technical efficiency, and highlight how the directional distance function becomes the corner-stone for the decomposition of overall economic (profit) efficiency under non-homotheticity.

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 in Section 3 the main characteristics, pros and cons, and relevant assumptions that need to be made to successfully undertake a study using DEA or SFA techniques. From a DEA perspective, we focus on the commonly used piecewise approximation of the production technology and comment on the main axioms regarding convexity, disposability, etc. We then show the programs allowing calculation of the directional and generalized distance functions. In the same vein, we present the main characteristics of the SFA when estimating these functions. These range from the choice of a flexible functional form, the imposition of the homogeneity properties that ensure a consistent characterization of the technology, the available estimation methods, distributional assumptions for the inefficiency component of the error term, etc. We discuss parameter restrictions that are required (or compulsory) by production theory, such as homogeneity conditions, versus alternative technological assumptions such as homotheticity, separability, etc.

Also, despite the fact that the gap between DEA and SFA is narrowing and that new proposals such as CNLS and StoNED are emerging, the truth is that most theoretical and empirical

issues are addressed from the perspective of one of the two approaches. Nevertheless, what ultimately matters is the reliability of results for managerial decision making and to better inform public policies from regulatory agencies and competition and anti-trust authorities—as discussed in Section 3.3. For this reason, we discuss in this paper common methodological and empirical topics aimed at improving use of both methods, and that scholars face regardless of the technique that is chosen to explain and measure economic performance. In this process there are critical modelling issues such as the available degrees of freedom for reliable statistical inference, and improving the discriminatory power of the methods. The need to reduce the dimensionality of the analysis may be present and it can be accomplished by using supervised and unsupervised multivariate reduction techniques, consistent aggregation, or the selection of the most relevant variables using statistical techniques. We also discuss the importance of direction selection in a section of its own. Once the restrictive input and output radial orientations have been definitely overcome by the more flexible directional and generalized specifications, the choice of direction becomes endogenous, and researchers must justify their preferred selection. This can be done in terms of the firms' economic objective, which is a natural choice in the framework that concerns this handbook, or alternative criteria based on the interpretation of the efficiency measures themselves−i.e., invariance to units of measurement, value of the score, etc. Recent contributions using data driven methods select local benchmarks that are similar to the firm under evaluation in terms of inputs and outputs mixes, or facing similar contextual (non-discretionary) variables, i.e., a similar productive and market environment.

As all these concerns are shared by both the non-parametric and parametric methods, we do not add to the almost endless debate on which approach is best, loosely based on their relative strengths and weaknesses, but advise the reader on the capabilities of each method to better address the existing empirical limitations and deal with research constraints. We may resort to two alternative narratives to exemplify the goal of this paper in helping scholars to choose the right model and techniques. Imagine first a study on the economic efficiency of firms operating in a regulated industry through a fix price regime (revenue-cap) or where the amount of output to be supplied is exogenous, and whose market structure accommodates a relatively small number of competitors. In this setting, given that outputs are non-discretionary, it seems logical to adopt a cost minimization approach, whose technological counterpart (by duality) is the input distance function, allowing to assess overall economic efficiency, and decompose it according to technical and allocative criteria. The fact that the industry accommodates a limited number of firms implies that there are not many observations, so resorting to SFA regression techniques with low degrees of freedom would be inadequate since results would lack statistical significance and based on weak inference. In this context, the use non-parametric DEA techniques will still yield individualized benchmarking results, useful for managerial analysis. This decision would be further reinforced from the empirical perspective if the data collection process is reliable with low measurement errors. Finally, if the relationship between inputs and outputs is complex, so external managerial knowledge is desirable, then including weight restrictions and environmental (non-discretionary) variables would be the preferred choice.

On the contrary, imagine now the case of a competitive industry, with a large number of competitors and where firms cannot exert market power. In this competitive environment, firms aim at maximizing profit by being technically efficient, and demanding and supplying the optimal amounts of inputs and outputs given their market prices. In this case the correct choice to assess their economic efficiency is to rely on the directional distance function measuring technical efficiency and, by duality, determine the allocative efficiency with respect to maximum profit. The

fact that there is a large number of observation allows taking full advantage of parametric regression analysis by estimating a quadratic function—amenable to the imposition of the traslation property, and capable of accommodating noise and measurement errors. This choice would be further supported if the required theoretical properties to be satisfied by the flexible functional form can be reasonably checked. Ultimately, this allows testing relevant characteristics of the production technology that cannot be easily identified relying on the DEA approach, such as homotheticity, returns to scale, marginal rates of substitution and transformation, etcetera.

We conclude emphasizing the relevance of the methods surveyed in this paper in unveiling the economic performance of firm in terms of technical and allocative (in)efficiencies, whose persistence and variability calls for further integration within the discipline of industrial organization. Efficiency and productivity analysis is now part of the toolbox in regulation and competition theory, providing the necessary analytical and quantitative results that allow the setting of firms' incentives in regulated industries (Agrell and Bogetoft, 2013), the evaluation of firms' market power through mark-ups (Abhiman and Kumbhakar, 2016), or the effects of mergers and acquisitions from the perspective of competition theory (Fiordelisi, 2009). Nevertheless, it is possible to think of additional fields where firms' heterogeneity in terms of their relative productivity is fundamental, as in the new trade models proposed by Melitz and Ottaviano (2008), where trade openness among countries triggers the Darwinian process of firm selection in domestic markets, with those situated in the lower tail of the (in)efficiency distribution exiting the industry. It is by now clear that the homogeneity associated to the canonical model of perfect competition is giving way to the reality associated to the indisputable evidence of inefficient behaviour. On these grounds, in terms of economic, technical and allocative fundamentals, the pieces of the inefficiency puzzle go towards explaining why firms deviate from best-practice operations and in this sense, make a valuable contribution to a wide range of research issues. As shown in the following contributions, 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.

Bibliography

Abhiman, D. and Kumbhakar, S.C. (2016) "Markup and Efficiency of Indian Banks: An Input Distance Function Approach", *Empirical Economics*, 51:4, 1689-1719.

Aczél, J. (1966), *Lectures on Functional Equations and their Applications,* New York: Academic Press.

Adler, N. and Golany, B. (2001) "Evaluation of Deregulated Airline Networks Using Data Envelopment Analysis Combined with Principal Component Analysis with an Application to Western Europe", *European Journal of Operational Research*, 132:2, 18-31.

Adler, N. and Golany, B. (2002) "Including Principal Component Weights to Improve discrimination in Data Envelopment Analysis", *Journal of the Operational Research Society*, 53, 985-991.

Adler, N. and Yazhemsky, E. (2010) "Improving Discrimination in Data Envelopment Analysis: PCA-DEA or Variable Reduction", *European Journal of Operational Research*, 202, 273-284.

Adragni, K.P. and Cook, D. (2009) "Sufficient Dimension Reduction and Prediction in Regression", *Philosophical Transactions of the Royal Society*, 367, 4385-4405.

Afriat, S. N. (1972) "Efficiency Estimation of Production Functions", *International Economic Review*, 13:3, 568-598.

Agee, M.D., Atkinson, S.E. and Crocker, T.D. (2012) "Child Maturation, Time-Invariant, and Time-Varying Inputs: Their Interaction in the Production of Child Human Capital", *Journal of Productivity Analysis*, 35, 29-44.

Ahmad, I. A. and Li, Q. (1997) ''Testing Independence by Nonparametric Kernel Method", *Statistics and Probability Letters*, 34, 201-210.

Ahn, S.C., Good, D.H. and Sickles, R.C. (2000) "Estimation of Long-Run Inefficiency Levels: a Dynamic Frontier Approach", Econometric *Reviews*, 19:4, 461-492.

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.

Akaike, H. (1973) "Information Theory and an Extension of the Maximum Likelihood Principle", In 2nd International Symposium on Information Theory, Petrov, B. and F. Csaki, T (Eds), 267-81, Budapest.

Allen, R., Athanassopoulos, A., Dyson, R.G. and Thanassoulis, E. (1997) "Weights Restrictions and Value Judgments in Data Envelopment Analysis: Evolution, Development and Future Directions", *Annals of Operations Research*, 73, 13-34.

Almanidis, Pavlos, Junhui Qian, and Robin C. Sickles. (2010) "Bounded Stochastic Frontiers with an Application to the US Banking Industry: 1984-2009." Rice University, (UnpublishedManuscript). (http://economics. rice. edu/WorkArea/DownloadAsset. aspx? id=-497) (2010).

Altunbas, Y., Evans, L. and Molyneux, P. (2001) "Bank Ownership and Efficiency", *Journal of Money, Credit and Banking*, 33:4, 926-954.

Altunbas, Y., Liu, M. H., Molyneux, P. and Seth, R. (2000) “Efficiency and Risk in Japanese Banking”, *Journal of Banking and Finance*, 24:10, 1605-1628.

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.

Álvarez, I., Barbero, J. and Zofio, J.L. (2016) "A Data Envelopment Analysis Toolbox for MATLAB," Working Papers in Economic Theory 2016/03, Department of Economics, Universidad Autónoma de Madrid, Spain, www.deatoolbox.com.

Amsler, C., Prokhorov, A. and Schmidt, P. (2014) "Using Copulas to Model Time Dependence in Stochastic Frontier Models”, *Econometric Reviews,* 33:5-6, 497-522.

Amsler, C., Prokhorov, A. and Schmidt, P. (2016) "Endogeneity in Stochastic Frontier Models”, *Journal of Econometrics,* 190:2, 280-288.

Anzanello, M.J. and Fogliatto, F.S. (2014) "A Review of Recent Variable Selection Methods in Industrial and Chemometrics Applications”, *European Journal of Industrial Engineering*, 8:5, 619-645.

Aparicio, J. and Zofío, J. L. (2017) “Revisiting the Decomposition of Cost Efficiency for Non-Homothetic Technologies: A Directional Distance Function Approach,” *Journal of Productivity Analysis, 48*:2-3, pp. 133-146.

Aparicio, J., Barbero, J., Kapelko, M., Pastor, J. T. and. Zofío, J.L. (2017) “Testing the Consistency and Feasibility of the Standard Malmquist-Luenberger Index: Environmental Productivity in World Air Emissions”, *Journal of Environmental Management*, 196, 148-160.

Aparicio, J., Borras, F., Pastor, J.T. and Zofío, J.L. (2016) “Loss Distance Functions and Profit Function: General Duality Results, in Aparicio, J., Lovell, C. A. K. and Pastor, J. T. (eds.), *Advances in Efficiency and Productivity*, Springer, 71-98.

Aparicio, J., Pastor, J. T. and Zofío, J. L. (2013) "On the Inconsistency of the Malmquist-Luenberger Index”, *European Journal of Operational Research*, 229:3, 738-742.

Aparicio, J., Pastor, J.T. and Zofío, J.L. (2015) “How to Properly Decompose Economic Efficiency Using Technical and Allocative Criteria with Non-Homothetic DEA Technologies,” *European Journal of Operational Research*, 240:3, 882-891.

Aparicio, J., Pastor, J.T. and Zofío, J.L. (2017) “Can Farrell's Allocative Efficiency be Generalized by the Directional Distance Function Approach?” *European Journal of Operational Research*, 257:1, 345-351.

Atkinson, S. E. and Tsionas, M. G. (2016) “Directional Distance Functions: Optimal Endogenous Directions”, *Journal of Econometrics*, 190:2, 301-314.

Balk, B. M. (1998) *Industrial Price, Quantity, and Productivity Indices. The Micro-Economic Theory and an Application,* Dordrecht: Kluwer Academic Publishers.

Balk, B. M. (2001) “Scale Efficiency and Productivity Change”, *Journal of Productivity Analysis,* 15, 159-183.

Balk, B. M. (2008) *Price and Quantity Index Numbers: Models for Measuring Aggregate Change and Difference*, New York: Cambridge University Press.

Balk, B. M. and Zofío, J.L. (2017) “The Many Decompositions of Total Factor Productivity Growth” Mimeo, Rotterdam School of Management, Erasmus University Rotterdam. The Netherlands.

Balk, B. M., (1997) "The decomposition of cost efficiency and the canonical form of cost function and cost share equations," *Economics Letters*, 55:1, 45-51.

Balk, B.M. (2017) "Profit-Oriented Productivity Change: A Comment", *Omega*, https://doi.org/10.1016/j.omega.2017.06.011.

Balk, B.M., Färe, R., and Grosskopf, S. (2004), “The theory of economic price and quantity indicators”, *Economic Theory*, 23:1, 149-164.

Balk, B.M., Färe, R., Grosskopf, S. and Margaritis, D. (2008) "Exact Relations Between Luenberger Productivity Indicators and Malmquist Productivity Indexes”, *Economic Theory*, 35:1, 187-190.

Bandyopadhyay, D. and Das, A. (2006) “On Measures of Technical Inefficiency and Production Uncertainty in Stochastic Frontier Production Model with Correlated Error Components”, *Journal of Productivity Analysis*, 26, 165-180.

Banker, R. D. and Natarajan, R. (2008) "Evaluating Contextual Variables Affecting Productivity Using Data Envelopment Analysis”, *Operations Research* 56:1, 48-58.

Banker, R. D. and Thrall R. M. (1992) “Estimation of Returns to Scale Using Data Envelopment Analysis.” *European Journal of Operational Research*, 62:1, 74-84.

Banker, R.D. and Morey, R. (1986) “Efficiency Analysis for Exogenously Fixed Inputs and Outputs”, *Operation Research*, 34:4, 513-521.

Banker, R.D., Charnes, A., Cooper, W.W., Swarts, J. and Thomas, D. (1989) “An Introduction to Data Envelopment Analysis with Some of its Models and their Uses”, *Research in Government and Nonprofit Accounting*, 5, 125-163.

Battese, G. E. and Coelli, T. J. (1992), “Frontier Production Functions, Technical Efficiency and Panel Data: With Application to Paddy Farmers in India”, *Journal of Productivity Analysis*, 3:1/2, 153-169.

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.

Battese, G.E., Rambaldi, A.N. and Wan, G.H. (1997) “A Stochastic Frontier Production Function with Flexible Risk Properties”, *Journal of Productivity Analysis*, 8, 269-280.

Battese, G.E., Rao, D.S.P., O’Donnell, C.J. (2004) “A Metafrontier Production Function for Estimation of Technical Efficiencies and Technology Gaps for Firms Operating under Different Technologies”, *Journal of Productivity Analysis*, 21:1, 1-103.

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.

Beard, T.R., Caudill, S.B. and Gropper, D.M. (1991) “Finite Mixture Estimation of Multiproduct Cost Function”, *The Review of Economics and Statistics*, 73:4, 654-664.

Bjureck, H. (1996), 'The Malmquist Total Factor Productivity Index', *Scandinavian Journal of Economics*, 98, 303-313.

Blackorby, C.and Diewert, W. E. (1979) "Expenditure Functions, Local Duality, and Second Order Approximations," *Econometrica*, 47:3, 579-602.

Bogetoft, P. and Otto, L. (2011) *Benchmarking with DEA, SFA, and R,* New York: Springer.

Bogetoft, P., Färe, R. and Obel, B. (2006) "Allocative Efficiency of Technically Inefficient Production Units", *European Journal of Operational Research,* 168, 450-462.

Boussemart, J.P., Briec W., Kerstens K. and Poutineau, J.C. (2003) "Luenberger and Malmquist Productivity Indices: Theoretical Comparisons and Empirical Illustration," *Bulletin of Economic Research,* 55:4,391-405.

Boussemart, J.P., Briec, W., Peypoch, N. and Tavéra, C. (2009) "α-Returns to Scale in Multi-Output Technologies" *European Journal of Operational Research,* 197:1, 332-339.

Bozdogan, H. (1987) "Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions", *Psychometrika*, 52(3), 345-370.

Bozdogan, H. (2000) "Akaike's information criterion and recent developments in information complexity", *Journal of Mathematical Psychology*, 44(1), 62-91.

Bravo-Ureta, B., Solís, D., Moreira-López, V., Maripani, J., Thiam, A. and Rivas, T. (2007) "Technical Efficiency in Farming: a Meta-Regression Analysis", *Journal of Productivity Analysis,* 27:1, 57-72.

Brons, M., Nijkamp,P., Pels, E. and Rietveld, P. (2005) "Efficiency of Urban Public Transit: a Meta Analysis", *Transportation,* 32:1, 1-21.

Brümmer, B., Glauben, T. and Thijssen, G. (2002) "Decomposition of Productivity Growth Using Distance Functions: The Case of Dairy Farms in Three European Countries." *American Journal of Agricultural Economics,* 84:3,628-644.

Bura, E. (2003) "Using Linear Smoothers to Assess the Structural Dimension of Regressions," *Statistica Sinica*, 13, 143-162.

Bura, E. and Cook, R.D. (2001) "Estimating the Structural Dimension of regressions Via Parametric Inverse Regression", *Journal of the Royal Statistical Society, Series B*, 63, 393-410.

Bura, E. and Yang, J. (2011) "Dimension Estimation in sufficient Dimension Reduction: A Unifying Approach", *Journal of Multivariate Analysis*, 102, 130-142.

Carree, M. A. (2002) "Technological inefficiency and the skewness of the error component in stochastic frontier analysis", *Economics Letters*, 77(1), 101-107.

Carta, A. and Steel, M.F.J. (2012) "Modelling Multi-Output Stochastic Frontiers Using Copulas", *Computational Statistics and Data Analysis*, 56:11, 3757-3773.

Caudill, S.B. and Ford, J.M. (1993) "Biases in Frontier Estimation Due to Heteroscedasticity", *Economic Letters,* 41, 17-20.

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.

Caves, D., Christensen, L. and Diewert, W.E. (1982) "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity*", Econometrica*, 50, 1393-1414.

Chamberlain, G. (1984) "Panel Data" in: Grilliches and Intriligator (eds.) *Handbook of Econometrics Vol. II*, Chapter 22, 1247-1318.

Chambers, R. (1983) "Scale and Productivity Measurement Under Risk", *American Economic Review,* 73, 802-805.

Chambers, R. (1998) *Applied Production Analysis: A Dual Approach*, New York: Cambridge University Press.

Chambers, R. (2002) "Exact Nonradial Input, Output, and Productivity Measurement", *Economic Theory*, 20:4, 751-765.

Chambers, R. and Quiggin, J. (2000) *Uncertainty, Production, Choice and Agency: The State-Contingent Approach,* New York: Cambridge University Press.

Chambers, R., Chung, Y. and Färe, R. (1996) "Benefit and Distance Functions", *Journal of Economic Theory,* 70, 407-419.

Chambers, R., Chung, Y. and Färe, R. (1998) "Profit, Directional Distance Functions and Nerlovian Efficiency", *Journal of Optimization Theory and Applications,* 95:2, 351-364.

Chambers, R.G., (1998) "Input and output indicators" in: Färe, R., Grosskopf, S., Russell, R.R. (eds.), *Index Numbers in Honour of Sten Malmquist*. Kluwer Academic Publishers, Boston, 241-272.

Charnes, A., Cooper, W.W. and Rhodes, E. (1981) "Evaluating Program and Managerial Efficiency: an Application of Data Envelopment Analysis to Program Follow Through", *Management Science,* 27:6, 668-697.

Chavas, J.P. (2008) "A Cost Approach to Economic Analysis under State-Contingent Production Uncertainty", American *Journal of Agricultural Economics*, 90:2, 435-446.

Chavas, J.P. and Cox, T. M. (1999) "A Generalized Distance Function and the Analysis of Production Efficiency", Southern *Economic Journal*, 66:2, 295-318.

Chen, C.M. (2009) "A Network-DEA Model with New Efficiency Measures to Incorporate the Dynamic Effect in Production Networks", *European Journal of Operational Research*, 194, 687-699.

Chen, C.M. (2014) "Evaluating eco-efficiency with data envelopment analysis: An analytical reexamination", *Annals of Operations Research*, 214, 49-71.

Chen, C.M. and Van Dalen, J. (2010) "Measuring Dynamic Efficiency: Theories and an Integrated Methodology", *European Journal of Operational Research*, 203, 749-760.

Chen, J., and Chen, Z. (2008) "Extended Bayesian information criteria for model selection with large model spaces" *Biometrika*, 95(3), 759-771.

Chen, J.X. (2012) “A Comment on DEA Efficiency Assessment Using Ideal and Anti-Ideal Decision Making Units”, *Applied Mathematics and Computation*, 219, 583-591.

Cherchye, L and Post, T. (2003) “Methodological Advances in DEA: A Survey and an Application for the Dutch Electricity Sector”, *Statistica Neerlandica*, 57:4, 410-438.

Chung, Y., Färe, R. and Grosskopf, S. (1997) “Productivity and Undesirable Outputs: A Directional Distance Function Approach”, *Journal of Environmental Management*, 51: 229- 240.

Coelli, T. J., Gautier, A., Perelman, S. and Saplacan- Pop, R. (2013) "Estimating the Cost of Improving Quality in Electricity Distribution: A Parametric Distance Function Approach”, *Energy Policy*, 53, 287-297.

Cook, R. D. and Weisberg, S. (1991) “Discussion of "Sliced Inverse Regression for Dimension Reduction", *Journal of the American Statistical Association*, 86, 28-33.

Cook, R.D. (2007) “Fisher Lecture: Dimension Reduction in Regression”, *Statistical Science*, 22:1, 1-26.

Cook, R.D. and Ni, L. (2005) "Sufficient Dimension Reduction Via Inverse Regression: A Minimum Discrepancy Approach”, *Journal of the American Statistical Association*, 100, 410-428.

Cook, W. D and Zhu, J. (2015) “DEA Cross Efficiency”, in J. Zhu (eds.), *Data Envelopment Analysis: A Handbook of Models and Methods,* New York: Springer.

Cooper, W.W., Seiford, L.M. and Tone, K. (2007) *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software,* Nueva Jersey: Princeton University Press.

Cordero, J. M., Santín, D. and Sicilia, G. (2015) “Testing the Accuracy of DEA Estimates Under Endogeneity through a Monte Carlo Simulation”, *European Journal of Operational Research,* 244:2, 511-518.

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.

Coyle, B. T. (2010), Production Economics, mimeo, University of Manitoba, Winnipeg, Canada.

Cuesta, R.A., Lovell, C.A.K. and Zofío, J.L. (2009) "Environmental efficiency measurement with translog distance functions: A parametric approach," *Ecological Economics*, 68:8/9, 2232-2242.

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.

Dakpo, K.H, Jeanneaux, P. Latruffe, L. (2016) “Modelling pollution-generating technologies in performance benchmarking: Recent developments, limits and future prospects in the nonparametric framework”, *European Journal of Operational Research*, 250:2, 347-359.

Daraio, C, and Simar, L. (2005) “Introducing Environmental Variables in Nonparametric Frontier Models: A Probabilistic Approach”, *Journal of Productivity Analysis* 24:1, 93-121.

Daraio, C. and Simar, L. (2016) “Efficiency and Benchmarking with Directional Distances: A Data-Driven Approach”, *Journal of the Operational Research Society*, 67:7, 928-944.

Das, A. and Kumbhakar, S.C. (2016) "Markup and Efficiency of Indian Banks: An Input Distance Function Approach”, Empirical *Economics,* 51:4, 1689-1719.

Debreu, G. (1951) “The Coefficient of Resource Utilization”, *Econometrica*, 19:3, 273-92.

Delis, M.D. and Tsionas, E.G. (2009) “The Joint Estimation of Bank-level Market Power and Efficiency”, *Journal of Banking and Finance,* 33, 1842-1850.

Diewert, W. E. (1971) "An Application of the Shephard Duality Theorem: A Generalized Leontief Production Function," *Journal of Political Economy*, 79, 461-507.

Diewert, W. E. (1976) “Exact and Superlative Index Numbers”, *Journal of Econometrics,* 4:2, 115-145.

Diewert, W. E. (1992) “Fisher Ideal Output, Input and Productivity Indexes Revisited”, *Journal of Productivity Analysis,* 3, 211-248.

Diewert, W. E. (2005)" Index Number Theory Using Differences Rather than Ratios", *The American Journal of Economics and Sociology*, 64:1, 311-360.

Dyson, R.G., Allen, R., Camanho, A.S., Podinovski, V.V., Sarrico, C.S. and Shale, E.A. (2001) “Pitfalls and Protocols in DEA”, *European Journal of Operational Research*, 132:2, 245-259.

Ellison, G. (1994) “Theories of Cartel Stability and the Joint Executive Committee”, *RandJournal of Economics*, 25:1, 37-57.

Emvalomatis, G. (2009) “Parametric Models for Dynamic Efficiency Measurement”, unpublished thesis.

Emvalomatis, G. (2012) “Adjustment and Unobserved Heterogeneity in Dynamic Stochastic Frontier Models”, *Journal of Productivity Analysis*, 37, 7-16.

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.

Fallah-Fini, S., Triantis, K., and Johnson, A.L. (2014) “Reviewing the literature on non-parametric dynamic efficiency measurement: state-of-the-art”, *Journal of Productivity Analysis*, 41, 51-67.

Fan, J. (1997) “Comments on «wavelets in statistics: A review”, by A. Antoniadis. *Statistical Methods and Applications*, 6(2), 131-138.

Fan, J. and Li R. (2001) “Variable Selection Via Nonconcave Penalized Likelihood and Its Oracle Properties”, *Journal of the American Statistical Association*, 96, 1348-1360.

Fan, J. and Lv, J. (2010) "A Selective Overview of Variable Selection in High Dimensional Feature Space”, *Statistica Sinica*, 20:1, 101-148.

Fan, J.; Li, R. (2006) Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery", In: Sanz-Sole, M.; Soria, J.; Varona, JL.; Verdera, J., editors. Proceedings of the International Congress of Mathematicians; p. 595-622.

Fan, Y., and Tang, C. Y. (2013) "Tuning parameter selection in high dimensional penalized likelihood", *Journal of the Royal Statistical Society: Series B* (Statistical Methodology), 75(3), 531-552.

Färe, R. and Grosskopf S. (2000b) "Theory and Applications of Directional Distance Functions", *Journal of Productivity Analysis*, 13:2, 93-103.

Färe, R. and Grosskopf, S. (2000a) "Notes on Some Inequalities in Economics", *Economic Theory,* 15:1, 227-233.

Färe, R. and Grosskopf, S. (2003) "Nonparametric Productivity Analysis With Undesirable Outputs: Comment", *American Journal of Agricultural Economics*, 85(4): 1070-1074.

Färe, R. and Grosskopf, S. (2004) "Modeling Undesirable Factors in Efficiency Evaluation: Comment" *European Journal of Operational Research*, 157:1, 242-245.

Färe, R., and Grosskopf, S. (1996) *Intertemporal Production Frontiers: With Dynamic DEA*, Boston: Kluwer Academic Publishers.

Färe, R., and Primont, D. (1995) *Multi-output production and duality: Theory and Applications,* Boston: Kluwer Academic Publishers.

Färe, R., Grosskopf, S. and Lovell, C. A. (2008) *Production Frontiers*, Cambridge University Press.

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. and Pasurka, C. (1986) "Effects on Relative Efficiency in Electric Power Generation due to Environmental Controls", *Resources and Energy,* 8, 167-184.

Färe, R., Grosskopf, S. and Pasurka, C.A. (2001) "Accounting for air pollution emissions in measures of state manufacturing productivity growth", *Journal of Regional Science*, 41:3, 381-409.

Färe, R., Grosskopf, S. and Pasurka, C.A. (2007) "Environmental production functions and environmental directional distance functions", *Energy*, 32, 1055-1066.

Färe, R., Grosskopf, S., Lindgren, B. and Roos, P. (1989, 1994) "Productivity Developments in Swedish Hospitals: A Malmquist Output Approach" in. Charnes, A., Cooper, W., Lewin, A. and Seiford, L. (eds.), Data *Envelopment Analysis: Theory, Methodology and Applications,* Dordrecht: Kluwer Academic Publishers.

Färe, R., Grosskopf, S., Lovell, C.A. and Pasurka, C. (1989) "Multilateral Productivity Comparisons When Some Outputs are Undesirable: A Nonparametric Approach", *The Review of Economics and Statistics*, 71:1, 90−98.

Färe, R., Grosskopf, S., Noh, D.W., Weber, W. (2005) "Characteristics of a Polluting Technology: Theory and Practice*", Journal of Econometrics*, 126, 469-492.

Farrell, M. (1957) "The Measurement of Productive Efficiency", Journal *of the Royal Statistical Society. Series A, General,* 120:3, 253-281.

Filippini, M. and Hunt, L.C. (2012) "US Residential Energy Demand and Energy Efficiency: A Stochastic Demand Frontier Approach", *Energy Economics*, 34:5, 1484-149.

Fiordelisi, F. (2009) *Mergers and Acquisitions in European Banking,* New York: Palgrave-MacMillan.

Fomby, T. B., Hill, R.C. and Johnson, S.R. (1978) "An Optimal Property of Principal Components in the Context of Restricted Least Squares", *Journal of the American Statistical Association* 73:361, 191-193.

Fonseca, J.R.S., Cardoso, M.G.M.S. (2007) "Mixture-Model Cluster Analysis Using Information Theoretical Criteria", *Intelligent Data Analysis*, 11:2, 155-173.

Førsund, F.R. (2009) "Good modelling of bad outputs: Pollution and multiple-output production, *International Review of Environmental and Resource Economics*, 3, 1-38.

Førsund, F.RA. and Kittelsen, S.A.C. (1998) "Productivity Development of Norwegian Electricity Distribution Utilities", *Resource and Energy Economics*, 20, 207-224.

Foster, D. and George, E. (1994) "The Risk Inflation Criterion for Multiple Regression", *Annals of Statistics*, 22, 1947-1975.

Frank, I.E. and Friedman, J.H. (1993) "A Statistical View of Some Chemometrics Regression Tools", *Technometrics*, 35, 109-148.

Fried, H., Lovell, C. A. and Shelton S. S. (2008) *The Measurement of Productive Efficiency and Productivity Growth*, New York: Oxford University Press.

Fried, H.O., Lovell, C.A., Schmidt, S.S. and Yaisawarng, S. (2002) "Accounting for Environmental Effects and Statistical Noise in Data Envelopment Analysis", *Journal of Productivity Analysis,* 17, 157-174.

Friedman, L. and Sinuany-Stern, Z. (1997) "Scaling units via the canonical correlation analysis in the DEA context", *European Journal of Operational Research*, 100:3, 25-43.

Fukuyama, H. and Weber, W.L. (2009) "A Directional Slacks-Based Measure of Technical Inefficiency", *Socio-Economic Planning Sciences*, 43:4, 274-287.

Fuss, M., McFadden, D. and Mundlak, Y. (1978) "A Survey of Functional Forms in the Economic Analysis of Production," in Fuss, M. and McFadden, D. (eds.), *Production Economic: A Dual Approach to Theory and Applications*, Amsterdam: North-Holland.

Gagnepain, P. and Ivaldi, M. (2002) "Stochastic Frontiers and Asymmetric Information Models", *Journal of Productivity Analysis,* 18:2, 145-159.

Galán, J. E. and Pollitt, M. G. (2014) "Inefficiency Persistence and Heterogeneity in Colombian Electricity Utilities", *Energy Economics*, 46, 31-44.

Galán, J. E., Veiga, H. andWiper, M. P. (2015) "Dynamic Effects in Inefficiency: Evidence from the Colombian Banking Sector", *European Journal of Operational Research*, 240:2, 562-571.

Gallant, A. R. (1984) "On the Bias in Flexible Functional Forms and an Essentially Unbiased Form." *Journal of Econometrics*, 15:211-45.

Gallant, A. R. (1984) "The Fourier Flexible Form," *American Journal of Agricultural Economics*, 66, 204-08.

Gaynor, M. and Anderson, G. F. (1995) "Uncertain Demand, the Structure of Hospital Costs, and the Cost of Empty Hospital Beds", *Journal of Health economics*, 14:3, 291-317.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014) *Bayesian data analysis* (Vol. 2). Boca Raton, FL: CRC press.

Giannakis, D., Jamasb, T. and Pollitt, M. (2005) "Benchmarking and Incentive Regulation of Quality of Service: An Application to the UK Electricity Distribution Networks", *Energy Policy,* 33:1, 2256-2271.

Golany, B. and Roll, Y. (1989) "An Application Procedure for DEA", *OMEGA*,17:3, 237-250.

Greene, W. (2002) *Econometric Analysis*, New York: Prentice Hall.

Greene, W. (2005) "Reconsidering Heterogeneity in Panel Data Estimators of the Stochastic Frontier Model", *Journal of Econometrics*, 126, 269-303.

Greene, W. (2008) "The Econometric Approach to Efficiency Analysis", in Fried, H., Lovell, C.A. and Schmidt, S.S. (eds.), *The Measurement of Productive Efficiency and Productivity Growth*, New York: Oxford University Press.

Greene, W. H. (2010) "A Stochastic Frontier Model with Correction for Sample Selection", *Journal of Productivity Analysis*, 34:1, 15-24.

Greene, W.H. (1990) "A Gamma-distributed Stochastic Frontier Model", *Journal of Econometrics*, 46:1-2, 141-164.

Griffin, R.C., Montgomery, J.M. and Rister, M.E. (1987) "Selecting Functional Form in Production Analysis," *Western Journal of Agricultural Economics*, 12, 216-227.

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.

Gropper, D.M., Caudill, S.B. and Beard. T.R. (1999)"Estimating Multiproduct Cost Functions Over Time Using a Mixture of Normals", *Journal of Productivity Analysis,* 11:3, 201-218.

Grosskopf, S. and Hayes, K. (1993) "Local Public Sector Bureaucrats and their Input Choices", *Journal of Urban Economics*, 33, 151-166.

Growitsch, C., Jamasb, T. and Wetzel, H. (2012) "Efficiency Effects of Observed and Unobserved Heterogeneity: Evidence from Norwegian Electricity Distribution Networks", *Energy Economics*, 34:2, 542-548.

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.

Guesmi, B. and Serra, T. (2015) "Can We Improve Farm Performance? The Determinants of Farm Technical and Environmental Efficiency", *Applied Economic Perspectives and Policy,* 37:4, 692-717.

Hailu, A. (2003) "Nonparametric Productivity Analysis with Undesirable Outputs: Reply," *American Journal of Agricultural Economics*, 85:4, 1075-1077.

Hailu, A., and Veeman T.S. (2001) "Non-parametric Productivity Analysis with Undesirable Outputs: An Application to the Canadian Pulp and Paper Industry" *American Journal of Agricultural Economics*, 83: 605-616.

Hall, P. and Li, K. C. (1993) "On Almost Linearity of Low Dimensional Projections from High Dimensional Data", *The Annals of Statistics*, 21, 867-889.

Hansen, C., J.B. McDonald and W.K. Newey (2010) "Instrumental Variables Estimation with Flexible Distributions," *Journal of Business and Economic Statistics*, 28, 13-25.

Hill, R.J. (2006) "Superlative index numbers: not all of them are super", *Journal of Econometrics*, 130: 25-43.

Hjalmarsson, L., Kumbhakar, S.C. and Heshmati, A. (1996) "DEA, DFA and SFA: a Comparison", *Journal of Productivity Analysis,* 7:2, 303-327.

Horrace, W. C. and Parmeter, C. F. (2014)"A Laplace stochastic frontier model" University of Miami Working Paper.

Huang, C.J., and Liu, J.T. (1994) "Estimation of a Non-neutral Stochastic Frontier Production Function", *Journal of Productivity Analysis*, 5:2, 171-180.

Huang, T. H., and Chen, Y. H. (2009) "A study on long-run inefficiency levels of a panel dynamic cost frontier under the framework of forward-looking rational expectations", *Journal of Banking and Finance*, 33:5, 842-849.

Huang, T.H., Chiang, D.L. and Chao, S.W. (2017) "A New Approach to Jointly Estimating the Lerner Index and Cost Efficiency for Multi-Output Banks under a Stochastic Meta-Frontier Framework", *Quarterly Review of Economics and Finance*, 65, 212-226.

Jamasb, T., Orea, L. and Pollitt, M. (2010) "Weather Factors and Performance of Network Utilities: A Methodology and Application to Electricity Distribution", EPRG Working Paper 1020, Cambridge Working Paper in Economics 1042.

Jenkins, L. and Anderson, M. (2003) "A Multivariate Statistical Approach to Reducing the Number of Variables in Data Envelopment Analysis", *European Journal of Operational Research*, 147, 51-61.

Johnson, A. L. and Kuosmanen, T. (2012) "One-stage and Two-stage DEA Estimation of the Effects of Contextual Variables", European *Journal of Operational Research,* 220:2, 559-570.

Johnson, A. L. and Kuosmanen, T. (2015) "An Introduction to CNLS and StoNED Methods for Efficiency Analysis: Economic Insights and Computational Aspects" in Ray S.C., Kumbhakar, S.C., and Dua, P. (eds.), *Benchmarking for Performance Evaluation: A Production Frontier Approach*, New Delhi: Springer.

Johnson, D. and McClelland, R. (1998) ''A General Dependence Test and Applications", Journal *of Applied Econometrics*, 13, 627-644.

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.

Juo, J-C., Fu, T-T., Yu, M-M.and Lin, Y-H. (2015) "Profit-oriented Productivity Change," *Omega*, 57, 176-187.

Just, R. E. and Pope, R. D. (1978) "Stochastic specification of production functions and economic implications", *Journal of Econometrics*, 7, 67-86.

Kallenberg, W.C. M. and Ledwina, T. (1999) ''Data-driven Rank Tests for Independence", *Journal of the American Statistical Association*, 94, 285-301.

Kapelko, M. (2017) "Dynamic versus Static Inefficiency Assessment of the Polish Meat-Processing Industry in the Aftermath of the European Union Integration and Financial Crisis" *Agribusiness*, 33:4, 505-521.

Kapelko, M. (2017) "Measuring productivity change accounting for adjustment costs: Evidence from the food industry in the European Union", *Annals of Operations Research*, forthcoming.

Kapelko, M. and Oude Lansink, A. (2017) "Dynamic multi-directional inefficiency analysis of European dairy manufacturing firms", *European Journal of Operational Research*, 257(1): 338-344.

Kapelko, M., Oude Lansink, A. and Stefanou, S. (2014) "Assessing Dynamic Inefficiency of the Spanish Construction Sector Pre- and Post-financial Crisis", *European Journal of Operational Research,* 237, 349-357.

Kapelko, M., Oude Lansink, A., and Stefanou, S. (2015) "Effect of food regulation on the Spanish food processing industry", PLoS ONE, 10(6): e0128217. doi:10.1371/journal.pone.0128217

Kapelko, M., Oude Lansink, A., Spiro, S.E. (2017) "Input-specific dynamic productivity change: Measurement and application to European dairy manufacturing firms", *Journal of Agricultural Economics*, 68 (2), 579-599.

Karakaplan, M.U. and Kutlu, L. (2015) "Handling Endogeneity in Stochastic Frontier Analysis". Unpublished manuscript.

Kerstens, K., Mounir, A. and Van de Woestyne, I. (2012) "Benchmarking Mean-Variance Portfolios Using a Shortage Function: the Choice of Direction Vector Affects Rankings!", *Journal of the Operational Research Society*, 63:9, 1199-1212.

Kim, M. (1986) "Banking Technology and the Existence of a Consistent Output Aggregate" , *Journal of Monetary Economics* 18:2 ,181-195.

Kittelsen, S.A.C. (1993) "Stepwise DEA: Choosing Variables for Measuring Technical Efficiency in Norwegian Electricity Distribution", Memorandum No. 06/93, Department of Economics, University of Oslo, Norway.

Koetter, M. and Poghosyan, T. (2009) "The Identification of Technology Regimes in Banking: Implications for The Market Power-Fragility Nexus", *Journal of Banking and Finance,* 33, 1413-1422.

Koetter, M., Kolari, J.W. and Spierdijk, L. (2012) "Enjoying the Quiet Life under Deregulation? Evidence From Adjusted Lerner Indices for U.S. Banks", *Review of Economics and Statistics*, 94:2, 462-480.

Konishi, S., and Kitagawa, G. (1996) “Generalised information criteria in model selection”, *Biometrika*, 83(4), 875-890.

Koop, R.J. (1981) “The measurement of productive efficiency: A reconsideration”, *Quarterly Journal of Economics*, 96, 477-503.

Koopmans, T. (1951) “An Analysis of Production as an Efficient Combination of Activities” in T. Koopmans (edt.) Activity Analysis of Production and Allocation, Cowles Commission for Research in Economics, Monograph. 13, John Wiley and Sons Inc., New York.

Kopp, R.J. and Diewert, W.E. (1982) “The Decomposition of Frontier Cost Function Deviations into Measures of Technical and Allocative Efficiency”, *Journal Econometrics,* 19, 319-331.

Krüger, J.J. (2012) "A Monte Carlo Study of Old and New Frontier Methods for Efficiency Measurement", *European Journal of Operational Research*, 222:1, 137-148.

Kumar, S. (2006) "Environmentally Sensitive Productivity Growth: A Global Analysis Using Malmquist-Luenberger Index", *Ecological Economics*, 56:2, 280-293.

Kumbhakar, S. and Lovell, C.A. (2000) *Stochastic Frontier Analysis*, Cambridge: Cambridge University Press.

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. (2010) "Efficiency and Productivity of World Health Systems: Where does your Country Stand?" *Applied Economics,* 42:13, 1641-1659.

Kumbhakar, S. C. (2012) "Specification and Estimation of Primal Production Models", *European Journal of Operational,* 217:3, 509-218.

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., Tsionas, E. G. and Sipiläinen, T. (2009) "Joint Estimation of Technology Choice and Technical Efficiency: an Application to Organic and Conventional Dairy Farming", *Journal of Productivity Analysis*, 31:2, 151-161.

Kumbhakar, S., L. Orea, A. Rodríguez-Álvarez and M. Tsionas (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.

Kumbhakar, S.C. (1991) “The Measurement and Decomposition of Cost- Inefficiency: the Translog Cost System” *Oxford Economic Papers,* 43:4, 667-683.

Kumbhakar, S.C. (1997) “Modeling Allocative Efficiency in a Translog Cost Function and Cost Share Equations: an Exact Relationship”, *Journal of Econometrics,* 76:1-2, 351-356.

Kumbhakar, S.C. (2011) "Estimation of Production Technology when the Objective is to Maximize Return to the Outlay", *European Journal of Operations Research*, 208:2, 170-176.

Kumbhakar, S.C. (2012) “Specification and Estimation of Primal Production Models”, *European Journal of Operational Research*, 217, 509-518.

Kumbhakar, S.C. and Efthymios, G. T. (2006) "Estimation of Stochastic Frontier Production Functions with Input-oriented Technical Efficiency", *Journal of Econometrics,* 133:1, 71-96.

Kumbhakar, S.C. and Tsionas, E.G. (2008) "Scale and Efficiency Measurement Using a Semiparametric Stochastic Frontier Model: Evidence from U.S. Commercial Banks", *Empirical Economics,* 34, 585-602.

Kumbhakar, S.C. and Tsionas, E.G. (2011) "Stochastic Error Specification in Primal and Dual Production Systems", *Journal of Applied Econometrics*, 26, 270-297.

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., Ghosh, S. and McGuckin, J.T. (1991) "A Generalized Production Frontier Approach for Estimating Determinants of Inefficiency in US Dairy Farms", *Journal of Business and Economic Statistics,* 9, 279-286.

Kumbhakar, S.C., Hung-Jen, W. and Horncastle, A.P. (2015) *"A Practitioner´s Guide to Stochastic Frontier Analysis Using Stata",* Cambridge University Press.

Kumbhakar, S.C., Park, B.U., Simar, L. and Tsionas, E.G. (2007) "Nonparametric Stochastic Frontiers: a Local Maximun Likelihood Approach", *Journal of Economics*, 137:1, 1-27.

Kumbhakar, S.C., Parmeter, C.F. and Tsionas, E.G. (2013) "A Zero Inefficiency Stochastic Frontier Model", *Journal of Econometrics*, 172:1, 66-76.

Kuosmanen, T. (2005) "Weak disposability in nonparametric production analysis with undesirable outputs", *American Journal of Agricultural Economics*, 87:4, 1077-1082.

Kuosmanen, T. and Kortelainen, M. (2005) "Measuring Eco-Efficiency of Production with Data Envelopment Analysis", *Journal of Industrial Ecology*, 9:4, 59-72.

Kuosmanen, T., and Kortelainen, M. (2005) "Measuring eco-efficiency of production with data envelopment analysis", *Journal of Industrial Ecology*, 9:4, 59-72.

Kuosmanen, T., Johnson, A.L., (2010) "Data envelopment analysis as nonparametric least squares regression", *Operations Research* 58 (1), 149-160.

Kutlu, L. (2010) "Battese-Coelli Estimator with Endogenous Regressors", *Economic Letters*, 109, 79-81.

Lai, H. P. (2013) "Estimation of the Threshold Stochastic Frontier Model in the Presence of an Endogenous Sample Split Variable", *Journal of Productivity Analysis*, 40:2, 227-237.

Lai, H.P. and Huang, C.J. (2010) "Likelihood Ratio Tests for ModelSelection of Stochastic Frontier Models", *Journal of Productivity Analysis*, 34:1, 3-13.

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.

Lai, H.P. and Huang, C.J. (2013) "Maximum Likelihood Estimation of Seemingly Unrelated Stochastic Frontier Regressions", *Journal of Productivity Analysis,* 40:1, 1-14.

Lau, L. J. (1972) "Profit Functions of Technologies with Multiple Inputs and Outputs", *Review of Economics and Statistic*s, 54, 281-289.

Lau, L.J. (1986) "Functional Forms in Econometric Model Building." *Handbook of Econometrics* 3, 1515-1566.

Lee, J.D., Park, J.B. and Kim, T.Y. (2002) "Estimation of the Shadow Prices of Pollutants with Production Environment Inefficiency Taken into Account: a Nonparametric Directional Distance Function Approach", *Journal of Environmental Management*, 63, 365-375.

Lee, L. (1983) "A test for distributional assumptions for the stochastic frontier function", *Journal of Econometrics,* 22:3, 245-267.

Lee, Y. H. and Schmidt, P. (1993) "A Production Frontier Model with Flexible Temporal Variation in Technical Efficiency. The measurement of productive efficiency: Techniques and applications, 237-255.

Leleu, H. (2013) "Shadow pricing of undesirable outputs in nonparametric analysis", *European Journal of Operational Research*, 231, 474-480.

Levinsohn, J. and Petrin, A. (2003) "Estimating Production Functions Using Inputs to Control for Unobservables", *Review of Economic Studies*,70, 317-341.

Levkoff, S.B. (2011), Ph.D. Chapter 3: Decomposing NOx and SO2 electric power plant emissions in a "By-production" framework: A nonparametric DEA study. In *Essays on efficiency measurement with environmental applications.* Riverside: Department of Economics, University of California.

Lewin, A.Y., Morey, R. C. and Cook, T. J. (1982) "Evaluating the Administrative Efficiency of Courts", *Omega*, 10:4, 401-411.

Li, K. (1991) "Sliced Inverse Regression for Dimension Reduction" (with discussion)*Journal of the American Statistical Association*, 86, 316-342.

Li, Q. (1996) "Estimating a Stochastic Production Frontier When the Adjusted Error Is Symmetric*", Economics Letters*, 52, 221-228.

Li, Q., Huang, C., Li, D. and Fu, T. (2002),"Semiparametric Smooth Coefficient Models", *Journal of Business and Economic Statistics,* 20:3, 412-422.

Liu, W., and Yang, Y. (2011) "Parametric or nonparametric? A parametricness index for model selection", *The Annals of Statistics*, 2074-2102.

Lovell, C. A., Rodríguez-Álvarez, A. and Wall, A. (2009) "The Effects of Stochastic Demand and Expense Preference Behavior on Public Hospital Costs and Excess Capacity", *Health Economics,* 18:2, 227-235.

Lucas, R. E. (1967) "Adjustment Costs and the Theory of Supply", *The Journal of Political Economy*, 75:4, 312-334.

Luenberger, D.G. (1992) "New Optimality Principles for Economic Efficiency and Equilibrium," *Journal of Optimization Theory and Applications*, 75:2, 221-264.

Madden, P. (1986) *Concavity and Optimization in Microeconomics*, New York, NY: Blackwell Publishers.

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.

Mas-Colell, A., Whinston, M.D. and Green, J.R. (1995) *Microeconomic Theory*, New York: Oxford University Press.

McElroy, M. (1987) "Additive General Error Models for Production, Cost, and Derived Demand or Share System", *Journal of Political Economy*, 95, 738-757.

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.

Melitz, M.J. and Ottaviano, G.I.P. (2008) "Market Size, Trade and Productivity", *Review of Economic Studies,* 75:1, 295-2316.

Mensah, Y.M. (1994) "A Simplification of the Kopp-Diewert Method of Decomposing Cost Efficiency and some Implications" *Journal of Econometrics,* 60, 113-144.

Mester, L. J. (1996) "A Study of Bank Efficiency Taking into Account Risk-Preferences", *Journal of Banking and Finance*, 20:6, 1025-1045.

Minkowski, H. (1911) "Theorie der Konvexen Körper," Gesammelte Abhandlungen II. Leipzig, Berlin: B.G. Teubner.

Mittelhammer, R.C., Judge, G.G. and Miller, D.J. (2000) *Econometric Foundations*, Cambridge University Press.

Morrison, C. J. (1992) "Unraveling the Productivity Growth Slowdown in the United States, Canada and Japan: the Effects of Subequilibrium, Sscale economies and Markups*", The Review of Economics and Statistics*, 381-393.

Mundlak, Y. (1961) "Empirical Production Function Free of Management Bias", *Journal of Farm Economics,* 43, 44-56.

Mundlak, Y. (1978) "On the Pooling of Time Series and Cross Section Data", *Econometrica,* 46, 69-85.

Muñiz, M.A. (2002) "Separating Managerial Inefficiency and External Conditions in Data Envelopment Analysis*", European Journal of Operational Research,* 143, 625-643.

Murty, S., Russell, R.R., and Levkoff, S.B. (2012) "On modeling pollution-generating technologies", *Journal of Environmental Economics and Management*, 64, 117-135.

Naik, P.A., Hagerty, M.R. and Tsai, C.L. (2000) "A New Dimension Reduction Approach for Data-rich Marketing Environments: Sliced Inverse Regression", *Journal of Marketing Research*, 37:1, 88-101.

Nemoto, J. and Goto, N. (1999) "Dynamic Data Envelopment Analysis: Modelling Intertemporal Behaviour of a Firm in the Presence of Productive Inefficiencies", *Economics Letters*, 64: 1, 51-56.

Nemoto, J. and Goto, N. (2003) "Measurement of Dynamic Efficiency in Production: An Application of Data Envelopment Analysis to Japanese Electric Utilities", *Journal of Productivity Analysis*, 19:2, 191-210.

Nguyen, N. B. (2010) "Estimation of Technical Efficiency in Stochastic Frontier Analysis, PhD thesis, Bowling Green.

Nieswand, M., Cullmann, A. and Neumann, A. (2009) "Overcoming Data Limitations in Nonparametric Benchmarking: Applying PCA-DEA to Natural Gas Transmission", DIW Discussion Papers, No. 962.

Norman M. and Stoker, B. (1991) *Data Envelopment Analysis: the Assessment of Performance*. New York: Wiley.

Nunamaker, T.R. (1985) "Using Data Envelopment Analysis to Measure the Efficiency of Non-Profit Organizations: a Critical Evaluation" *Managerial and Decision Economics,* 6:1, 50-58.

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.

O'Donnell, C.J., Rao, D.S.P. and Battese, G. (2008) "Metafrontier Frameworks for the Study of Firm-Level Efficiencies and Technology Ratios", *Empirical Economics*, 34:2, 231-55.

Odeck, J. and Brathen, S. (2012) "A Meta-Analisys of DEA and SFA Studies of the Technical Efficiency of Seaports: a Comparison of fixed and Random-Effects Regression Models", *Transportation Research Part A: Policy and Practice,* 46:10, 1574-1585.

Oh, D.H. and Lee, J.D. (2010) "A Metafrontier Approach for Measuring Malmquist Productivity Index, *Empirical Economics,* 38:1, 47-64.

Olley, G.S. and Pakes, A. (1996) "The Dynamics of Productivity in the Telecommunications Equipment Industry", *Econometrica*, 64, 1263-1297.

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 (2017) "A Parametric Approach to Estimating Eco-Efficiency", *Journal of Agricultural Economics,* 68:3, 901-907.

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., Growitsch, C. and Jamasb, J. (2015) "Using Supervised Environmental Composites in Production and Efficiency Analyses: an Application to Norwegian Electricity Networks", *Competition and Regulation in Network Industries,* 16:3, 260-288.

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.

Orea, L., Roibás, D. and Wall, A. (2004) "Choosing the Technical Efficiency Orientation to Analyze Firms Technology: a Model Selection Test Approach", *Journal of Productivity Analysis,* 22:1-2, 51-71.

Orme, C. and Smith, P. (1996) "The Potential for Endogeneity Bias in Data Envelopment Analysis", *Journal of the Operational Research Society*, 47:1, 73-83.

Orzechowski, W. (1977) "Economic Models of Bureaucracy: Survey, Extensions and Evidence", in Borcherding, T.E. (eds.), *Budgets and Bureaucrats: The Sources of Government Growth,* Durham, NC: Duke University Press.

Oude Lansink, A., Stefanou, S.E., and Serra, T. (2015) "Primal and dual dynamic Luenberger productivity indicators", *European Journal of Operational Research*, 241, 555-563.

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.

Parmeter, C.F., Wang, H.J. and Kumbhakar, S.C. (2017) "Nonparametric Estimation of the Determinants of Inefficiency", *Journal of Productivity Analysis*, 47:3, 205-211.

Pasiouras, P. (2008) "Estimating the Technical and Scale Efficiency of Greek Commercial Banks: The Impact of Credit Risk, Off-Balance Sheet Activities and International Operations", *Research in International Business and Finance*, 22:3, 301-349.

Pastor, J. M. and Serrano, L. (2005) "Efficiency, Endogenous and Exogenous Credit Risk in the Banking Systems of the Euro Area", Applied *Financial Economics*, 15:9, 631-649.

Pastor, J. T. and Aparicio, J. (2010) "Distance Functions and Efficiency Measurement," *Indian Economic Review*, 45:2, 193-231.

Pastor, J.T. Ruiz, J.L. and Sirvent, I. (2002) "A Statistical Test for Nested Radial DEA Models" *Operations Research,* 50:4, 728-735.

Pastor, J.T., Aparicio, J., Alcaraz, J., Vidal, F. and Pastor, D. (2016) "The Reverse Directional Distance Function", in Aparicio, J., Lovell, C. A. and Pastor, J. T. (eds.), *Advances in Efficiency and Productivity*, Springer.

Peyrache, A. and Coelli, T. (2009) "Testing Procedures for Detection of Linear Dependencies in Efficiency Models", European *Journal of Operational Research*, 198:2, 647-654.

Peyrache, A. and Daraio, C. (2012) "Empirical Tools to Assess the Sensitivity of Directional Distance Functions to Direction Selection," *Applied Economics*, 44:8, 933-943.

Picazo-Tadeo, A., Beltrán-Esteve, M. and Gómez-Limón, J. (2012) "Assessing Farming Eco-Efficiency: a Data Envelopment Analysis Approach", *European Journal of Operational Research*, 220:3, 798-809.

Picazo-Tadeo, A., Reig-Martínez, E. and Gómez-Limón, J. (2011) "Assessing Farming Eco-Efficiency: a Data Envelopment Analysis Approach", *Journal of Environmental Management*, 92:4, 1154-1164.

Pittman, R.W. (1983) "Multilateral Productivity Comparisons with Undesirable Outputs", *The Economic Journal*, 93,883-91.

Podinovski, V. V. (2015) "DEA Models with Production Trade-offs and Weight restrictions", in Zhu, J. (eds.), *Data Envelopment Analysis: A Handbook of Models and Methods,* New York: Springer.

Podinovski,V.V., and Kuosmanen, T. (2011) “Modelling weak disposability in data envelopment analysis under relaxed convexity assumptions”, *European Journal of Operational Research*, 211, 577-585.

Pope, R.D. and Chavas, J.P. (1994) "Cost Functions under Production Uncertainty", *American Journal of Agricultural Economics*, 76, 114-127.

Rasmussen, S. (2004) "Optimizing Production under Uncertainty: Generalisation of the State-Contingent Approach and Comparison of Methods for Empirical Application", Paper presented at the 2004 Asia-Pacific Productivity Conference, Brisbane, 14-16 July.

Ray, S.C. (1988.) "Data Envelopment Analysis, Nondiscretionary Inputs and Efficiency: an Alternative Interpretation", Socio- *Economic Planning Science*, 22:4, 167-176.

Ray, S.C. (2004) “*Data Envelopment Analysis. Theory and Techniques for Economics and Operations Research”,* Cambridge University Press.

Reifschneider, D. and Stevenson, R. (1991) "Systematic Departures from the Frontier: A Framework for the Analysis of Firm Inefficiency", *International Economic Review* 32: 3,715-723.

Richmond, J. (1974)"Estimating the Efficiency of Production", *International Economic Review*, 15:2, 515- 521.

Robinson, P. M. (1988) "Root-NConsistent Semiparametric Regression", *Econometrica* 56:4, 931-954.

Rockafellar, R.T. (1972), *Convex Analysis*, Princeton, New Jersey: Princeton University Press. Second Printing.

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.

Rodríguez-Álvarez, A., Roibás, D. and Wall, A. (2012) "Reserve Capacity of Public and Private Hospitals in Response to Demand Uncertainty", *Health Economics*, 21:7, 839-851.

Ruggiero, J. (1996) "On the Measurement of Technical Efficiency in the Public Sector", *European Journal of Operational Research,* 90, 553-565.

Ruggiero, J. (1998) “Non-Discretionary Inputs in Data Envelopment Analysis”, *European Journal of Operational Research*, 111:3, 461-459.

Rungsuriyawiboon, S. and Stefanou, S. E. (2007) "Dynamic Efficiency Estimation: An Application to U.S. Electric Utilities", *Journal of Business and Economic Statistics*, 25:2, 226-238.

Samuelson, P. A. (1979) "Paul Douglas's Measurement of Production Functions and Marginal Productivities," *Journal of Political Economy*, 87:5, 923-939.

Santín, D. and Sicilia, G. (2017) “Dealing with Endogeneity in Data Envelopment Analysis Applications”, *Expert Systems with Applications*, 68, 173-184.

Sato, R. (1977) “Homothetic and Non-Homothetic CES Production Functions,” *The American Economic Review,* 67:4, 559-569.

Saunders, H. (2008)"Fuel Conserving (and using) Production Functions", *Energy Economics*, 30:5, 2184-2235.

Schmidt, P. and Lovell, C.A.K. (1979), "Estimating technical and allocative inefficiency relative to stochastic production and cost frontiers", *Journal of Econometrics,* 9:3, 343-366.

Schwarz, G. (1978) "Estimating the Dimension of a Model", *Annals of Statistics*, 6, 461-644.

Seiford, L.M. and Zhu, J. (2002) "Modeling Undesirable Factors in Efficiency Evaluation" *European Journal of Operational Research*, 142:1, 16-20.

Seiford, L.M. and Zhu, J. (2005) "A Response to Comments on Modeling Undesirable Factors in Efficiency Evaluation" *European Journal of Operational Research*, 161:2, 579-581.

Sengupta, J. K. (1995) *Dynamics of Data Envelopment Analysis*, Dordrecht: Kluwer Academic Publishers.

Sengupta, J.K., 1990. "Tests of efficiency in data envelopment analysis", *Computers and Operations Research*, 17:2, 123-132.

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.

Serra, T., Stefanou, S., and Oude Lansink, A. (2010) "A Dynamic Dual Model under State-Contingent Production Uncertainty", *European Review of Agricultural Economics*, 37:3, 293-312.

Sexton, T. R., Silkman, R. H. and Hogan, A. J. (1986) "Data Envelopment Analysis: Critique and Extension" in Silkman, R. H. (eds.), *New Directions for Program Evaluation*, San Francisco: Jossey-Bass.

Shephard, R.W. (1953) *Cost and Production Functions*, Princeton, New Jersey: Princeton University Press.

Shephard, R.W. (1970) *Theory of Cost and Production Functions*, Princeton, New Jersey: Princeton University Press.

Silva, E. and Oude Lansink, A. (2009) "Dynamic Efficiency Measurement: a directional Distance Function Approach. Unpublished manuscript. Wageningen: Wageningen University.

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.

Silva, E. and Stefanou, S. (2003) "Nonparametric Dynamic Production Analysis and the Theory of Cost", *Journal of Productivity Analysis,* 19:1, 5-32.

Silva, E. and Stefanou, S. (2007) "Nonparametric Dynamic Efficiency Measurement: Theory and Application*", American Journal of Agricultural Economics,* 89:2, 398-419.

Silva, E., Oude Lansink, A., and Stefanou, S.E. (2015) "The adjustment-cost model of the firm: duality and productive efficiency", *International Journal of Production Economics*, 168, 245-256.

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.

Simar, L. and Wilson, P. W. (2010)"Inferences from Cross-Sectional, Stochastic Frontier Models", *Econometric Reviews,* 29, 62-98.

Simar, L. and Wilson, P.W. (2007) "Estimation and Inference in Two-Stage, Semi-Parametric Models of Production Processes", *Journal of Econometrics*, 136:1, 31-64.

Sinuany-Stern, Z. and Friedman, L. (1998) "DEA and the discriminant analysis of ratios for ranking units", *European Journal of Operational Research*, 111:3, 470-478.

Sinuany-Stern, Z. Mehrez, A. and Barboy, A. (1994) "Academic Departments Efficiency Via DEA" *Computers and Operations Research,* 21:5, 543-556.

Skevas, T., Lansink, A. O. and Stefanou, S. E. (2012)"Measuring Technical Efficiency in the Presence of Pesticide Spillovers and Production Uncertainty: The Case of Dutch Arable Farms", *European Journal of Operational Research*, 223:2, 550-559.

Smith, M.D.(2008), "Stochastic Frontier Models with Dependent Error Components", *Econometrics Journal*, 11, 172-192.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002) "Bayesian measures of model complexity and fit", *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(4), 583-639.

Stevenson, R. E. (1980) "Likelihood Functions for Generalized Stochastic Frontier Estimation." *Journal of Econometrics,* 13:1, 57-66.

Sueyoshi, T., and Sekitani, K. (2005) "Returns to scale in dynamic DEA", *European Journal of Operational Research*, 161, 536-544.

Sun, K. and Kumbhakar, S.C. (2013)"Semiparametric Smooth-Coefficient Stochastic Frontier Model", *Economics Letters,* 120, 305-309.

Thompson, G. D. (1988) "Choice of Flexible Functional Forms: Review and Appraisal," *Western Journal of Agricultural Economics*, 13:2, 169-183.

Thompson, R.G., Singleton, F.D., Thrall, R.M. and Smith, B.A. (1986) "Comparative Site Evaluations for Locating a High-Energy Physics Lab in Texas," *Interfaces* 16, 35-49.

Tian, L., Cai, T., Goetghebeur, E., and Wei, L. J. (2007) "Model evaluation based on the sampling distribution of estimated absolute prediction error", *Biometrika*, 94(2), 297-311.

Tibshirani R. (1996) "Regression Shrinkage and Selection Via the LASSO", *Journal of the Royal Statistical Society. Series* B, 58:1, 267-288.

Tone, K., and Tsutsui, M. (2010) "Dynamic DEA: A slacks-based measure approach", OMEGA, 38, 145-156.

Tone, K., and Tsutsui, M. (2014) "Dynamic DEA with network structure: a slacks-based measure approach", OMEGA, 42, 124-131.

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. (2013) "GMM Estimation of Stochastic Frontier Model with Endogenous Regressors", *Economics Letters,* 118, 233-236.

Tran, K.C. and Tsionas, E.G. (2015) "Endogeneity in Stochastic Frontier Models: Copula Approach without External Instruments", *Economics Letters*, 133, 85-88.

Treadway, A. B. (1969) "On Rational Entrepreneurial Behaviour and the Demand for Investment", *Review of Economic Studies*, 36:2, 227-239.

Treadway, A. B. (1970) "Adjustment Costs and Variable Inputs in the Theory of the Competitive Firm", *Review of Economic Studies*, 2:4, 329-347

Tsionas, E.G. (2006) "Inference in Dynamic Stochastic Frontier Models", *Journal of Applied Econometrics,* 21:5, 669-676.

Tsionas, E.G., Kumbhakar, S.C. and Malikov. E. (2015) "Estimation of Input Distance Functions: A System Approach", *American Journal of Agricultural Economics*, 97:5, 1478-1493.

Tyteca, D. (1996) "On the measurement of the environmental performance of firms—A literature review and a productive efficiency perspective", *Journal of Environmental Management*, 46, 281-308.

Tyteca, D. (1997), "Linear Programming Models for the Measurement of Environmental Performance of Firms—Concepts and Empirical Results", *Journal of Productivity Analysis*, 8:2, 183-197.

Ueda, T. and Hoshiai, Y. (1997) "Application of Principal Component Analysis for Parsimonious Summarization of DEA Inputs and/or Outputs", *Journal of the Operational Research Society of Japan,* 40, 466-478.

Vardanyan, M. and Noh, D.W. (2006) "Approximating Pollution Abatement Costs Via Alternative Specifications of a Multi-Output Production Technology: a Case of the U.S. Electric Utility Industry", *Journal of Environmental Management, 80:2,* 177-190.

Velicer, W.F. and Jackson, D.N. (1990) "Component Analysis versus Common Factor Analysis: Some Issues in Selecting an Appropriate Procedure", *Multivariate Behavioral Research*, 25:1, 89-95.

Verbeek, M. (2008) *A Guide to Modern Econometrics*, John Wiley and Sons, Ltd.

Verschelde, M., Dumont, M., Rayp, G. and Merlevede, B. (2016) " Semiparametric Stochastic Metafrontier Effiency of European Manufacturing Firms", *Journal of Productivity Analisys,* 45:1, 53-69.

Vuong, Q. H. (1989) ''Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses'', *Econometrica*, 57:2, 307-333.

Wagner, J.M. and Shimshak, D.G. (2007) "Stepwise Selection of Variables in Data Envelopment Analysis: Procedures and Managerial Perspectives", *European Journal of Operational Research*, 180, 57-67.

Wang, H., Li, R., and Tsai, C. L. (2007) "Tuning parameter selectors for the smoothly clipped absolute deviation method", *Biometrika*, 94(3), 553-568.

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 (2009) “On the Distribution of Estimated Technical Efficiency in Stochastic Frontier Models”, *Journal of Econometrics*, 148, 36-45.

Wilson, P. W. (2003) "Testing Independence in Models of Productive Efficiency", *Journal of Productivity Analysis,* 20:3, 361-390.

Wilson, P.W. (2008) “FEAR: A Software Package for Frontier Efficiency Analysis with R”, *Socio-Economic Planning Sciences*, 42:4, 247-254.

Wooldridge, J. (2009)"On Estimating Firm-Level Production Functions Using Proxy Variables to Control for Unobservable", *Economics Letters*, 104, 112-114.

Xia, Y., Tong, H., Li, W. K. and Zhu, L.X. (2002) "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society, Ser. B*, 64,363-410.

Yang, H.L., and Pollitt, M. (2009) “Incorporating both undesirable outputs and uncontrollable variables into DEA: The performance of Chinese coal-fired power plants”, *European Journal of Operational Research*, 197,1095-1105.

Yu, W., Jamasb, T. and Pollitt, M. (2009) “Does Weather Explain Cost and Quality Performance? An Analysis of UK Electricity Distribution Companies”, *Energy Policy*, 37:11, 4177-4188.

Żak-Szatkowska, M., and Bogdan, M. (2011) “Modified versions of the Bayesian information criterion for sparse generalized linear models”, *Computational Statistics and Data Analysis*, 55(11), 2908-2924.

Zellner, A. and Revankar, N.S. (1969) "Generalized Production Functions", *Review of Economics and Statistics,* 36:2, 241-250.

Zellner, A. Kmenta, J. and Dreze, J. (1966) Specification and Estimation of Cobb-Douglas Production Function Models", *Econometrica*, 34, 784-795.

Zheng, J. X. (1997) ‘‘A Consistent Specification Test of Independence.’’ *Nonparametric Statistics*, 7, 297-306.

Zhu, J. (1998) “Data Envelopment Analysis vs. Principle Component Analysis: An Illustrative Study of Economic Performance of Chinese Cities”, *European Journal of Operational Research*, 11, 50-61.

Zhu, L., Baiqi M. and Peng, H. (2006) "On Sliced Inverse Regression with High-Dimensional Covariates", *Journal of the American Statistical Association*, 101: 474, 630-643.

Zieschang, K.D. (1983) “A Note on the Decomposition of Cost Efficiency into Technical and Allocative Components”, *Journal of Econometrics,* 23:3, 401-405.

Zofio, J. L. and Prieto, A. M. (2001) "Environmental Efficiency and Regulatory Standards: the Case of CO2 Emissions from OECD Industries", *Resource and Energy Economics*, 23:1, 63-83.

Zofío, J. L. and Prieto, A. M. (2006) “Return to Dollar, Generalized Distance Function and the Fisher Productivity Index”, *Spanish Economic Review*, 8:2, 113-138.

Zofío, J. L., Pastor, J. and Aparicio, J. (2013) "The Directional Profit Efficiency Measure: On Why Profit Inefficiency is either Technical or Allocative", *Journal of Productivity Analysis,* 40:3, 257-266.