INFERRING REAL DEA-ESTIMATORS FROM SMALL FRONTIER DISPLACEMENTS: A MONTE CARLO STUDY

VICTOR MAIA SENNA DELGADO

Abstract. The convergence of technical efficiency of $(p + q)$ estimators remains an open question in the field of Nonparametric frontier models. Various techniques, such as Free Disposal Hull (FDH) and Data Envelopment Analysis (DEA), are employed to delineate the efficiency frontier. Park et al. (2000) [1] and Daouia et al. (2010) [2] established that the FDH estimator converges to a Weibull distribution through the application of Extreme Value Theory (EVT). The $(p + q)$ dimensional convergence to the case of $(p + q) \leqslant 4$ is assured. The most challenging case involves this greater dimensional convergence for DEA estimators. In 2008, Kneip, Simar, and Wilson [3] demonstrated the convergence of the DEA-VRS using bootstrap-based algorithms. In this study, we introduce an alternative approach to constructing intervals for DEA-estimators and subsequently investigate its convergence (or not) to real Data Generating Process Models. Monte-Carlo simulations indicate that our first results align with existing literature.

1. Introduction

Productivity and efficiency analysis involves measuring how effectively inputs are converted into outputs and identifying factors that influence this conversion. The study of production frontiers has a long-standing tradition in economics and operational research, [4] and [5], and could be applicable in many fields. It is used for many purposes: (i) identifying best practices among the most efficient units, which can serve as *benchmarks* for others; (ii) to understand the technology underlying a sector or industry; (iii) study time-series productivity trends and (iv) to guide governmental mission in providing public goods [6] (among other uses). Following the nomenclature proposed by [7] let the production set Ψ be:

$$\Psi = \{(x, y) \in \mathbb{R}_+^{p+q} \mid x \text{ can produce } y\}. \tag{1.1}$$

Where x is an input-vector defined in $\mathbb{R}_+^p$. The length of the input-vector is p, and y is the output vector in $\mathbb{R}_+^q$. So the production set have q different outputs. The simple case $p = 1$ and $q = 1$ is going to be discussed in the simplest model of this article, section 2. A generalization $p \times q$ for $p \geqslant 1$ or $q \geqslant 1$ will be discussed in the section 3 and as indication for further work in this topic.

The production set have a boundary defined by the maximum output obtained to a given level of inputs. Or, by other way, the minimum input given a certain level of outputs. This

2020 *Mathematics Subject Classification.* 60E05, 62G05, 62G20, 90C05.

JEL-Codes: C44, C61, D20.

Key words and phrases. Frontier analysis, Nonparametric Frontiers, Extreme Value Theory (EVT), Data Envelopment Analysis (DEA), Monte Carlo Method.

could be obtained proportionally to all inputs (outputs), which is called *radial* efficiency or for at least one input (output) changing, forming the *non-radial* efficiency. The frontier of the production set is defined by:

$$\Psi^{\partial} = \{(x, y) \in \Psi \mid (\theta x, \theta^{-1} y) \notin \Psi \text{ for any } \theta < 1\}. \tag{1.2}$$

Typically, θ is a scalar and the equation (1.2) immediately gives a radial efficiency. On the definition above, any $\theta > 1$ is outside the production set Ψ (see figure 1). Also, the θ could stand for a vector and the $\theta \cdot x$ is a scalar-product. It will be outside the production set for any value in θ greater than one for a particular unit i.[1] The point (x, y) (or unit of analysis) are recurrently called as *Decision Making Units* (or simply as DMUs) by the literature.

Note that including its border, and considering only the input-efficiency, we can rewrite the equation (1.1) more directly (the same can be done for output-efficiency):

$$\Psi = \{(x, y) \in \mathbb{R}_{+}^{p+q} \mid \theta \in (0, 1]\}. \tag{1.3}$$

We are going to append three more assumptions to the production set:

(1) **Free Disposal**: For (x, y) inside the production set and (x', y') such that $x' \geqslant x$ and $y' \geqslant y$, the observation (x', y') is also inside the production set (i.e. $(x', y') \in \Psi$).
(2) **No Free Lunch**: When $x = \mathbf{0}$ then y must be a vector of zeroes. This assumption says that it is impossible to produce something out of nothing. Formally, if $x = \mathbf{0}$ and $y \gg 0$ (i.e. y is a strict positive vector, not including any zeroes), then $(x, y) \notin \Psi$. Note that this does not exclude the contrary: $y = \mathbf{0}$ and $x \gg 0$ is possible and means that production set require some investment.
(3) **Convexity**: The production set is convex, that is, for any (x, y) and $(x', y') \in \Psi$ we have: $\alpha(x, y) + (1 - \alpha)(x', y') \in \Psi$ for all $\alpha \in [0, 1]$.

Those are fair economic assumptions. All of them may be relaxed if the purpose requires. It is less common to dispense with (1) and (2), and it is very common (for some purposes) not to apply (3) or maybe add some supposition of the returns to scale. The Free Disposal Hull (FDH) method [9] to obtain Ψ^{∂} doesn't require the latter assumption. For the methods to be developed in the next section, it is convenient to keep all three, (1) to (3), although perhaps in further developments of this research we should check again if they are all attended or check if we are not making (implicitly) some extra assumption.

The figure 1 presents the **theoretical** *Production Set* in observance of those three economic assumptions above. The shaded region is the set Ψ, its border, where $\Psi^{\partial} = \{\Psi \mid (\theta = 1)\}$ is generally included, as the set is said to be **compact** (*closed* and *bounded*). We also have the complementary set Ψ^{c} or, alternatively, points $\notin \Psi$. Note that the origin $(0, 0)$ is not a must in this set, but the *no free lunch* exclude any vertical intercept.

[1] For outputs, θ^{-1} is a vector with inverse values, say λ, then $\lambda \cdot y$ is the product scalar **outside** the frontier to any value in λ less than 1, or $\frac{1}{\theta} < 1$. In a similar equation, Simar and Wilson [7] use a more general vector called γ instead θ or λ, because it could be used in many different methods for measuring a distance from a particular point (x, y) to the frontier: hyperbolic, directional distances and other methods [8] (section 2).

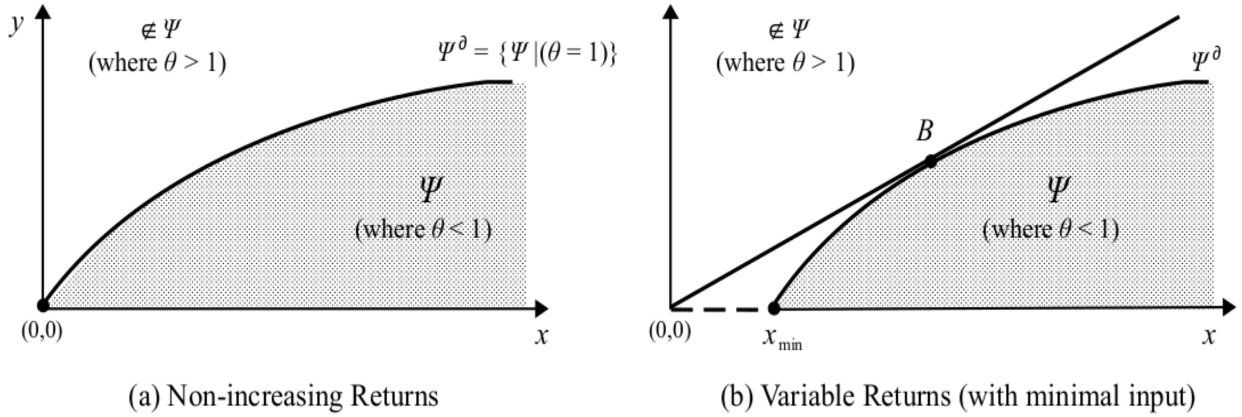


FIGURE 1. The production set, the frontier Ψ^∂ and its complementary space.

The set represented by figure 1(a) have *convexity* and *non-increasing* returns to scale (NIRS) and the set of figure 1(b) have all three assumptions attended but Variable Returns to Scale (VRS), before point B it presents increasing returns and after that point it presents decreasing returns to scale. Exactly in the point, we have Constant Returns to Scale (CRS). In x, the segment from 0 to $x_{\min}$ could be in the set or not, if it is, the possibility of no production is maintained but the property (3) of convexity is not, if it is not the set remains convex. In the latter case, it could be said that Ψ presents the *sunk cost* property, as $x_{\min}$ is irreversible.

The *Data Envelopment Analysis* method with Constant Returns to Scale (CRS), known as DEA-CRS [10], and its variant with Variable Returns to Scale (VRS), known as DEA-VRS [11], along with the previously mentioned FHD [9] are three most popular nonparametric methods for constructing envelopment estimators of Ψ^∂. We are going to call here all these estimators as DEA-estimators or DEA-type estimators. Since the boundary function is not known to empirical researchers, these estimation methods are essential.

As documented by professors Moradi-Motlagh and Emrouznejad [12], the last two decades in the area of nonparametric boundaries have seen an increasing use of statistical methods to obtain the properties of θ, largely due to two highly influential papers by Simar and Wilson [13, 14]. These papers compelling made the argument for the proper way to perform the bootstrap estimators, and addressed the correction and interpretation of the bias $(\hat{\theta}_i - \theta_i)$ for any DMU i or for all DMUs in the observational set $i = \{1, \ldots, n\}$, let's call it I whenever necessary.

Prior to this literature, the *deterministic* nature of the nonparametric approach was often highlighted as a positive feature. This emphasis reinforced the importance of high-quality data measurement to prevent the "garbage-in, garbage-out" phenomenon.

This perspective was challenged by the recognition that productive and especially social data are prone to mismeasurements. And those are difficult to minimize for both x and y. It introduces two simultaneous sources of uncertainty. It is important to note that, for parametric estimators like *Ordinary Least Squares* (OLS), errors in x do not hold as much prominence.

However, in nonparametric frontiers, these errors can be crucial in determining whether an observation is deemed efficient or not. Furthermore, the source of the error – whether it arises from x or y – is generally unknown.

In addition to [7, 12], a valuable resource for the latest advances is the book by professors Robin Sickles and Valentin Zelenyuk (2019, chapters 9 and 10) [15]. We have adapted the Motlagh's et al. (2022) diagram to encompass this important developments. It provides a schematic representation of the key developments in the literature, including references to significant papers. It is beyond our scope here to make a detailed literature review of each of the texts, but we advise the interested reader to look for the main texts related to each of the diagram boxes.

The robust statistics approach is characteristic of stochastic processes for obtaining the production set border Ψ^{∂}. The boxes highlighted in red in the diagram are the most closely related to this work. Although, we are not going to make borders of the order-m type o frontiers.

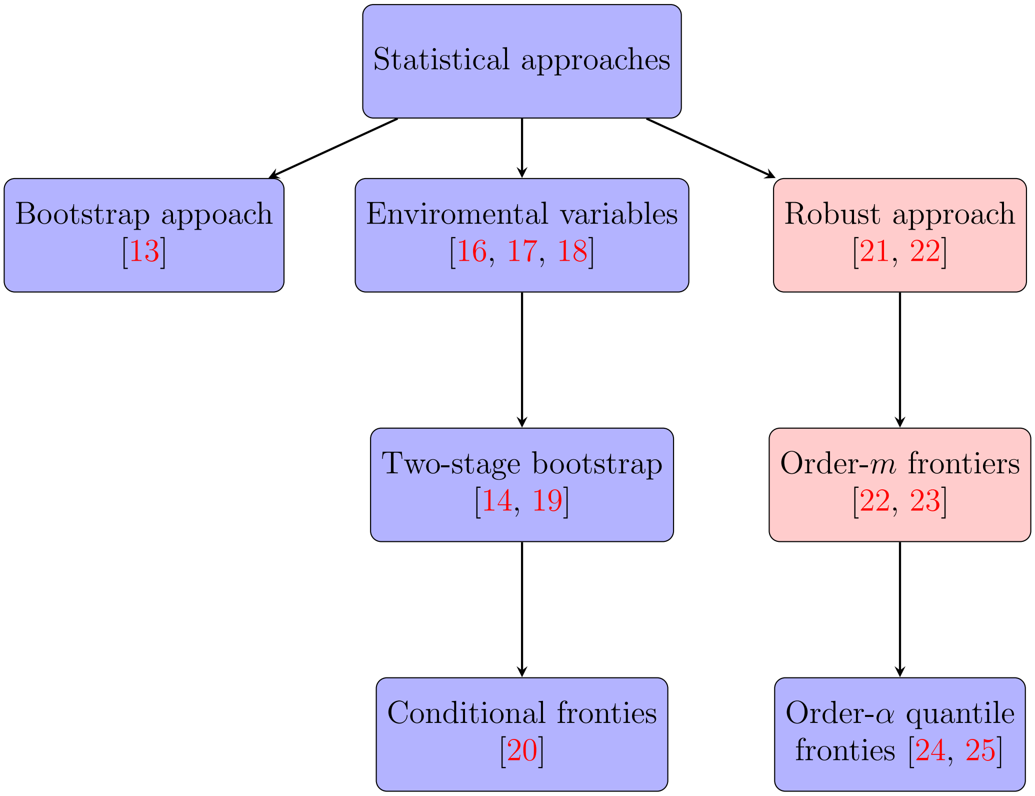

Statistical approaches
Bootstrap appoach [13]
Enviromental variables [16, 17, 18]
Robust approach [21, 22]
Two-stage bootstrap [14, 19]
Order-m frontiers [22, 23]
Conditional fronties [20]
Order-α quantile fronties [24, 25]

Diagram 1. Statistical approaches in the literature in Nonparametric Frontier Analysis.

Our main purpose is to construct a robust stochastic approach related with the Extreme Value Theory (EVT). We assume that one of the advantages of such a method may be to obtain a higher frontier that "envelops" all the data points, and that in the near future (with more advances to be discussed along this text and also in the conclusion part). Another objective is to perform statistical inference with Monte-Carlo construct data. We highlight that all these developments can be of use by applied research.

In the next section (section 2), we present the construction of our approach, giving more functional details. Section 3 generalizes this perspective in a formal way to $(p+q) \geqslant 3$ dimensions and pointing to asymptotics properties. In the section [4], we present Monte Carlo simulations using familiar data generating process data. The section 5 concludes our perspective and points some of the shortcomings or limitations of our work, pointing to future research and further developments.

2. Small Frontier Displacements

2.1. **Underlying Data Assumptions.** First, we present the intuition of the frontier method proposed here with a graphical exposition in $\mathbb{R}^2$. Before entering the details of Figure 2, let's represent $\mathcal{V}(\Psi)$ to denote the convex cone of Ψ. This is previous to observed data to be discussed:

$$\mathcal{V}(\Psi)^{\partial} = \left\{(x,y) \in \mathcal{V}(\Psi) \mid (\theta x, \theta^{-1}y) \notin \mathcal{V}(\Psi) \text{ for any } \theta < 1\right\} \tag{2.4}$$

If the $\mathcal{V}(\Psi)$ is exactly equal to Ψ then this is valid for its border and then the function is said to have *globally* Constant Returns to Scale (CRS). Alternatively, if $\Psi \subseteq \mathcal{V}(\Psi)$ with $\Psi \cap \mathcal{V}(\Psi) = (x,y)$, a single point, then the Ψ^{∂} have *locally* non-increasing returns to scale (NIRS).[2] The empirical construction of $\mathcal{V}$ and Ψ requires some data realization: $\mathcal{X}_n = \{(X_i, Y_i)\}_{i=1}^{n}$, which is, for each randomization, generating the number sequence. This is the Data Generating Process to be described also in section 4. More precisely:

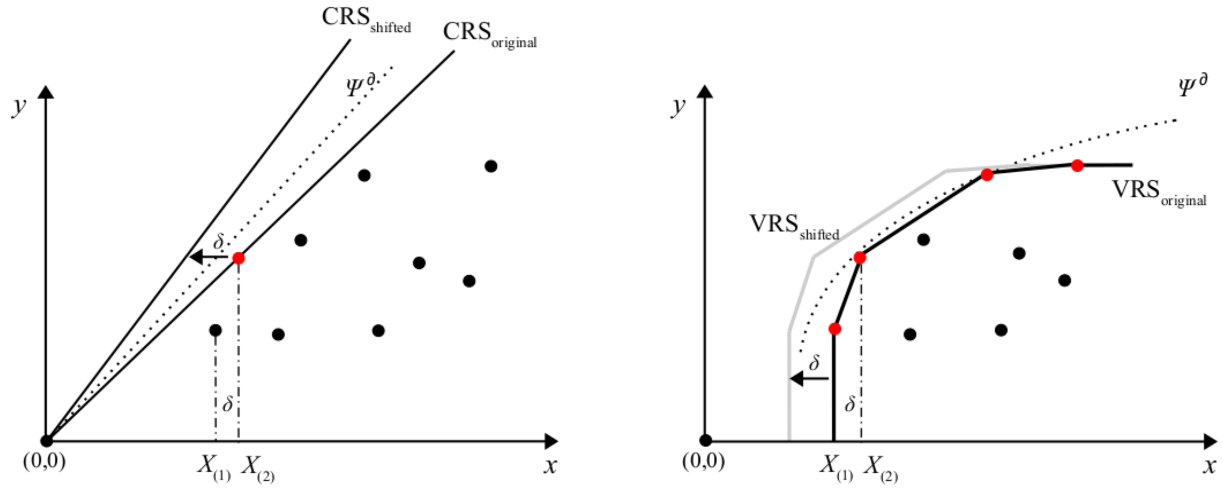
$$(X,Y) = \{(x_i, y_i) \mid i = \{1, 2, \dots, n\}\} \tag{2.5}$$

Where X and Y may be interpreted as random variables linked to the DGP $\mathcal{X}_n$. It is usually required that the underlying DGP function assumptions follow the (1)-(3) assumptions of the previous section (or (1) and (2) to a non-convex estimation for the set).[3] Additionally, borrowing from Park, Simar and Weiner's approach [1] to enunciate the above construction in the probabilistic setup, we make the following assumptions:

(1) The observations (X_i, Y_i) are independent and identically distributed random variables (*iid*) with common density f defined on the support Ψ.

[2] For an observational process, the intersection set $\Psi \cap \mathcal{V}(\Psi)$ is going to have always $(p+q)-1$ dimensions. Note that the NIRS depends on one production set as the figure 1(a) graph, not the 1(b). The proof of these affirmations could be made with the hyperplane separator theorem. The converse type of frontier, the always increasing returns to scale (IRS), could also be made, but not both, NIRS and IRS, simultaneously, which is impossible.

[3] It is important to add the detail that the DEA-VRS estimator do not require (NIRS) but does require (3). Usually the subset where $f(X,Y)$ "starts" to be positive, the beginning of the observational production set is increasing in the VRS model.



(a) Small displacement CRS

(b) Small displacement VRS

FIGURE 2. Small Displacement of the original frontier with CRS and VRS

(2) Regularity conditions on the density f near the frontier. At the frontier the density f is positive, i.e. $f_{(x,y)} = f(\theta x, y) > 0$, and sequentially Lipschitz-continuous.[4]

(3) Differentiability and concavity. Let $\varphi(X)$ be a function $\mathbb{R}^p \mapsto \mathbb{R}$ of the production set border Ψ^∂ already described. Then this assumption imposes $\frac{\partial}{\partial x}\varphi(x,y) > 0$ and $\frac{\partial^2}{\partial x^2}\varphi(x,y) < 0$ for any input x^k in the production set and in the vicinity of a particular (x, y) of interest.[5]

Park at al. [1] also introduce the assumptions of monotonicity and convexity to production function. With (1)-(3) of the Section 1 along with (1) to (3) mentioned above, Park et al. proved that the asymptotic distribution of any $\hat{\theta}$ converges to a Weibull distribution function. Those six assumptions may be sufficient (broader sense) to provide an adequate description of the DGP.[6]

2.2. **Graphical and algorithmic construction.** Given a set of random variables (X, Y) provided by a specific realization of a generating process, in order to construct an efficiency frontier, we have selected a slightly alternative approach to obtain robust estimators of the efficiency.

[4] For all sequences (x^n, y^n) in Ψ converging to a particular point $(\theta x_0, y_0)$ it follows that:

$$|f(x^n, y^n) - f(\theta x_0, y_0)| \leqslant C_1 \|(x^n, y^n) - (\theta x_0, y_0)\| \quad \text{for some} \quad C_1 > 0.$$

[5] The notation becomes more complex when incorporating the conditioning details of Park et al. [1].

[6] We didn't find much recent literature about *sufficient statistic* of FDH or DEA estimators, but Bugetoft (1993, p. 260) [26] relating to Andersen and Petersen [27] super-efficiency models and concerned with agency-problems of frontier estimation wrote: "The construction of more aggregated, *sufficient statistics* depends on the specific class of possible frontiers [Ψ] considered and the prior beliefs [...] on [Ψ] assumed. Furthermore, optimal contracts generally underutilize even minimal sufficient statistics, but exactly how depends on the details of the contracting problem." The emphasis on *sufficient statistics* is ours.

Our approach is somewhat similar to that of Hall et al. (1998) [8], who used a stochastic frontier method, albeit nonparametric. This approach closely aligns with the perspective of Bădin and Simar (2009) [23], although it is simpler and was developed independently prior to becoming aware of their work. It is also inspired by the work of [13, 22] and [3], building robustness through bootstrap processes (see sections 3 and 4 of this text).[7]

Figure 2 presents an $\mathbb{R}^2$ example where $p = 1$ and $q = 1$. The panel (a) presents the "Displacement method" described by the steps of construction below (2.2.1) for a CRS estimator construction, and the panel (b) presents it for the VRS and FDH methods. Little more generalization is presented in 2.2.2. Those two approaches resonates with the Extreme Value Theory (EVT), which we will explore further.

2.2.1. *Displacement method.*

- 1. For an observed dataset, (X, Y) compute the DEA-type frontier and the efficiency for each point i, giving $\hat{\theta}_i$.
- 2. From the (X, Y) obtain $X_{(1)} = \min\{X_i\}$ and $Y_{(n)} = \max\{Y_i\}$, first order statistics.
- 3. Repeat the step 1 for the second order statistic of X. So, if $X_{(1)}$ is the j^{th} observation, obtain $X_{(2)} = \min\{X_i\}_{i\neq j, i\in I}$ and compute: $\delta = |X_{(1)} - X_{(2)}|$. Do the same for the $(n-1)$-order statistic of Y: $Y_{(n-1)} = \max\{Y_i\}_{k\neq i, i\in I}$, where the k^{th} observation is the $Y_{(n)}$. Distance $\delta' = |Y_{(n-1)} - Y_{(n)}|$.
- 4. Subtract (or sum) δ from a subset of (X, Y) in the frontier (red dots indicated in Figure 2), let's call it $(X, Y)^*$. This is the *shift* indicated by the arrows. For an *input-shift*, the new frontier will be formed by the $(X - |\delta|, Y)^*$ set (it is $(X, Y + |\delta|)^*$ for an *output-shift*). Obtain the efficiency of each point in the shifted frontier: $\theta_{s,i}$.[8]
- 5. Compute the average between $\hat{\theta}$ and θ_s for each point i in the data set. This is the frontier displacement efficiency θ_d.

The Figure 2 also indicates that the real frontier border Ψ^∂ could be between the observed and shifted frontier. The Figure 3, in section 4, shows that construction with a Monte Carlo procedure realization. In fact, any directional *shift* is possible, and this is the perspective for the general approach below.

One of the possible extensions to $\mathbb{R}^{p+q}$ occurs, making a δ for each input variable X (or output variable Y). Making the same process for all variables is equivalently to contract (or expand) it radially. If the researcher have more information about the data, she/he could manage to construct a δ-vector with the directional values for the frontier. Even, if there is sufficient reason, it can at the same time contract X and expand Y, moving the boundary directionally.

2.2.2. *More general approach.*

- 1. Compute the DEA-type frontier.
- 2. For each input p in the production set do the step 2 and 3 of 2.2.1.[9]
- 3. From the step above, compute the vector $\boldsymbol{\delta}$.

[7] In fact, to add more detail in advance, we didn't present the bootstrap in this paper, only Monte Carlo simulations, but it will be ready for forthcoming works.

[8] The θ_s stands for efficiency of frontier shifted.

[9] The same can be done by changing X to Y.

- 4. For each corresponding δ_p in $\boldsymbol{\delta}$, obtain $(X - \delta I, Y)^*$ and compute the *shifted*-frontier.
- 5. Compute the average between $\hat{\theta}$ and θ_s for each point i in the data set obtaining θ_d.

We are going to *loop* the 2.2.1 five-step processes from 25 up to 2,000 Monte Carlo process simulations in section 4 (the 2.2.2 more general setting we are going to develop in further works). We should say that some restrictions to δ or data could be imposed. For example, for a DGP close enough to the origin, or particularly close to the vertical axis and with a large variance. It all depends on how realistic or not these assumptions are for an observed data set, or for the stronger requirement of knowing the *Data Generating Process* of all observations.

2.3. **Some considerations about the approach.** Various other construction methods can be thought of, nonparametric estimation of the boundary. Alternative convex forms might also prove useful. The FDH process achieves multiple convex corners, suggesting that a piecewise convex approach could offer a more flexible form to infer the true frontier. Also, higher and lower orders could be used for X and Y respectively, getting more information on these statistics can be useful for even greater robustness.[10] The question is “how far inner we want to go?” or “Is it recommended covering all data in this way?” In the conclusions, we will return to these questions, but we advance here that the answer is that “we do not know”. More studies are needed to know if there is a recommendation on it.

We can say that the data do not exist from “above” the frontier (the complementary Ψ space), however, for the researcher in front of a DGP that she does not know also don't know when the space of the non-border is going to give some information, i.e. $1 - F_{X|Y}(\theta x|y) > 0$ (see section 3 for details). This displacement approach can be exactly good for testing hypothesis. Hyperbolic techniques could also be thought, or some semicircular approach for the data.

3. Formalizing the method to $\mathbb{R}^{(p+q)}$ (Some useful theorems)

As pointed in [1], one of the main interest in productivity and efficiency analysis is the production or technology set Ψ, i.e., a set of technically possible pair of input $\mathbf{x} \in \mathbb{R}^p_+$ and output $\mathbf{y} \in \mathbb{R}^q_+$. It is essential to estimate the production set from a sample drawn from the observed population.

First, let's consider the following way to define the efficiency estimator:

$$\theta(x, y) = \inf \{\theta \mid F(\theta x|y) > 0\} \tag{3.6}$$

Where $F(x|y)$ is the *marginal* cumulative density function of x conditional on y. Fortunately, this function has an empirical countable version:

$$\widehat{F}(x|x) = \frac{\sum_n \mathbb{1}(X_i \leqslant x, Y_i \geqslant y)}{\sum_n \mathbb{1}(Y_i \geqslant y)} \tag{3.7}$$

[10] See Bădin and Simar (2009) [23].

Where $\mathbb{1}$ is the indicator function (the counter) and we are computing this in our n observation sample. Also for a multidimensional perspective, [17] showed that the FDH estimator, in fact, is:

$$\widehat{\theta}_{FDH}(x,y) = \min_{i, Y_i \geqslant y} \left\{ \max_{p} \left(\frac{X_i^p}{x^p} \right) \right\} \tag{3.8}$$

Where p is in the set of *all* possible inputs and j is a particular input in that set, i.e. $j = \{1, \ldots, p\}$. Finally, considering that μ^{p+q} is a $(p+q)$-dimensional mean to a multidimensional DGP, then we have the Park et al. (2000) [1] theorem:

Theorem 3.1. *With (1)-(2) assumptions on support Ψ (section 1) and (1)-(3) assumptions on the DGP (section 2), then for all $z > 0$ and $i \in I$ we have:*

$$P\Big[n^{\frac{1}{p+q}} (\theta_i - \widehat{\theta}_i) \leqslant z \Big] = 1 - e^{-(\mu z)^{p+q}} + o(1).$$

And the important following corollary, guarantee under Lindeberg-Feller Central Limit Theorem of finite-variance:

Corollary 3.2. *Asymptotically:*

$$n^{\frac{1}{(p+q)}} \Big(\widehat{\theta}_{FDH}(x,y) - \theta(x,y) \Big) \xrightarrow{\mathcal{L}} W(\mu^{p+q}, p+q)$$

Where $W(\mu^{p+q}, p+q)$ is the Weibull distribution with μ^{p+q} and $(p+q)$ as parameters. Using some properties of Weibull distribution with, $\mu = 1$ the authors also prove the following theorem (note that r must be greater to the negative of $(p+q)$, which appears as the $p+q+r > 0$ condition):

Theorem 3.3. *With (1)-(3) assumptions on the DGP (section 2), then:*

$$\mathbb{E}[(\theta - \widehat{\theta}_{FDH})^r] = c_r \frac{1}{(\mu^{(p+q)})^r} \cdot \frac{1}{n^{\frac{r}{p+q}}} + o\left(\frac{1}{n^{\frac{r}{p+q}}} \right)$$

Where

$$c_r = \Gamma\left(\frac{p+q+r}{p+q} \right)$$

Following a Gamma distribution with p, q and r as parameters (r could be equal 1 or 2 for most typical cases). The proofs for these both theorems (and the asymptotic corollary) are advanced. We are developing an original proof for these to be included in the appendices. The focus is on the asymptotic simulations that follow in section 4. One of the conclusion's recommendations is to improve on this topic. However, it should be noted that the results are important for the inferences to be made.

4. Monte Carlo Procedures

In this section, we are going to describe the realized Monte Carlo procedures by three parts. We focused mainly in the input-oriented approach, but everything could be transposed to output-oriented as well. Four *DGP* were specified (so-called **models**):

1. Linear: $Y = \beta X + \varepsilon$;
2. Linear with increasing returns: $Y = \alpha_0 + \beta X + \varepsilon$;
3. Strict concave: $Y = X^\beta + \varepsilon$;
4. Strict concave with initial costs, i.e. increasing returns to scale: $Y = \alpha_1 + X^\beta + \varepsilon$, also only valid in $\mathbb{R}_+$.

Values for α_0, α_1 and β are -10, -1 and 0.5, respectively. The domain for X is defined in the interval 0 to 100, i.e. $X \in [0, 100]$. and that ε is always subtracted from equations (1) to (4) from the definition above. So error distribution assumptions are particular for this kind of simulation, $\varepsilon \in [0, -Y^*]$, where Y^* is the theoretical maximum giving the GDP process. It gives error as uniform with varying interval conditional on X. This increases error variance as X increases, another error specifications are possible, however, allowing errors to vary more widely is beneficial for examining certain properties of indicators (see [25]).

4.1. **Monte Carlo for efficiencies.** We explored seven different ranges for n size, from 25 up to 2,000: $\{25,\ 50,\ 100,\ 200,\ 500,\ 1,000,\ 2,000\}$ and four different DEA estimators: CRS, NIRS, VRS and FDH (all discussed in section 1 and 2).

Figure 3 shows the simple example of four DEA estimating procedures to $(Y = \beta X - |\varepsilon|)$ model with $n = 25$. It summarizes our general procedure described in section 2. The *gray* color frontiers are the frontier with displacements, the small shifts. Note that even for DRS, VRS and FDH, sometimes the displaced frontier will be left and above the real frontier (the dotted line). Remembering the (1.2) definition, the shifted border for an input-oriented frontier will be:

$$\Psi^{\partial}_{shift} = \{(x - |\delta|, y) \in \Psi \mid (\theta(x - |\delta|), y) \notin \Psi \text{ for any } \theta < 1\} \tag{4.9}$$

And considering the efficiency of any DMU i comparing with this border $(\Psi^{\partial}_{shift})$, let's call it $\theta_{s,i}$ and the original efficiency $(\widehat{\theta}_i)$, we have the efficiency computed by displacement method $(\theta_{d,i})$ as:

$$\theta_{d_i} = \frac{\theta_{s,i} + \widehat{\theta}_i}{2} \tag{4.10}$$

It is expected that the real frontier will be in the interval between the computed $\widehat{\theta}$ and the shifted one θ_s. Specifying our main research problem as a **hypothesis test**, we have:

- Null hypothesis (H_0): $\theta_d = \theta$
- Alternative (H_1): $\theta_d \neq \theta$

As X is defined in the interval $[0, 100]$, let's make it vary by a uniform distribution, $X \sim U(0, 100)$. Considering Y^* as the theoretical maximum giving by a *DGP* process, $Y^* = \beta X^*$ (or any equation of the beginning of this section), we have Y^* determined.

Then, for each X_i realization, the absolute errors are computed by a uniform distribution $Y_i \sim U(0, Y_i^*)$.

As suggested by [2, 25], this particular kind of Y distribution is interesting to explore the possibility of envelopment estimators converge or not.[11] In the conclusion of this section, we discuss the implications of our specifications.

In a Monte Carlo procedure, we know the real θ and we are going to discuss the distribution of θ_d. In the section 3, we discussed the theoretical Weibull distribution encountered by literature [1], in the next subsection we are going to present the empirical distribution of the proposed θ_d and its p-values comparing with a t-distribution and Weibull distribution.[12] Checking for bias (or not) the Hypothesis test will become:

- Null hypothesis (H_0): $\mathbb{E}(\hat{\theta}_d - \theta) = 0$
- Alternative (H_1): $\mathbb{E}(\hat{\theta}_d - \theta) \neq 0$

Specifically, it is very important to consider that the DEA estimators are inherently biased and that traditional Central Limit Theorems (CLT) are not applicable to these estimators. Kneip et al. (2015) [29] propose certain specifications and literature searching to mitigate these issues.

While we have attempted to incorporate these suggestions, several constraints have limited the final results presented. Consequently, further developments are necessary, which will be discussed in more detail in the conclusions section 5. Nonetheless, some of the results here obtained points to possible interesting properties of the small displacement methods.

Figure 4 presents the illustration of the Monte Carlo, increasing the n-size. Particularly, in panel (d), we have that the difference between VRS and FDH frontier to real frontier. It is almost imperceptible for an error small enough.

4.2. **Estimating the Bias and making the inference.** Table 1 make the important case for the input-orientation of the center locus of the graph. It is possible to see the convergence of DEA-VRS estimator (with displacement) for all four models. That converge is slow, as known in DEA-type estimators.

Figure 5 shows all 4 models with three DEA-estimators (*with displacement*) of the frontier (INRS was suppressed because it's very close to VRS). Panel (a) shows a good convergence of all models, particularly for CRS, which has the correct returns to scale specification (**model 1**). In panel (b), VRS and FDH do great job for **model 2**, but CRS, despite the rapid convergence, presents large bias because returns to scale mispecification (it is the price that the CRS specification charges for having to pass through the origin). Almost same case for panel (c) and **model 3** $Y = X^\beta$ and panel (d) which converges slowly and CRS do well but with negative bias (most likely because although the border does not pass through the origin), and VRS and FDH presents a steady convergence.

The empirical standard deviation is presented as $S(\theta_d)$. It shows an empirical constancy of the variation, except for the first three rows, where some oscillation is expected and row

[11] There are more extreme specifications for errors, with the distribution mass far from frontier or from more sparse errors (*noisy data*). For expansion of these techniques, see Daouia et al. (2020) [28].

[12] We are still leaving the t-distribution for test for normality and check some properties of proposed method. Contrary to traditional methods, DEA estimators have very asymmetrical distribution of the bias statistic.

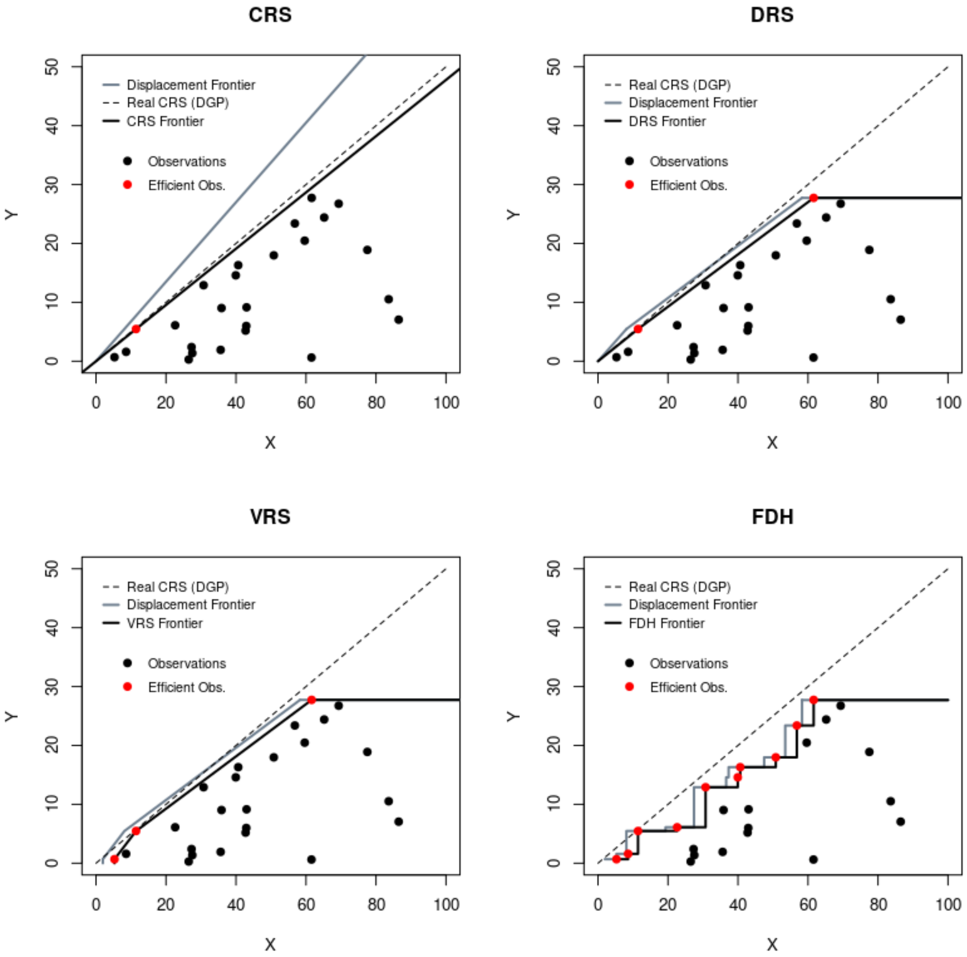


FIGURE 3. Monte Carlo simulations for DGP Model 1, $n = 25$, and four DEA-type estimators.

4 and 5 for **model 1**, the standard deviation computed is around 0.2 or 0.3. This slow decreasing in the values of the standard errors is expected in DEA models.

Considering the definitions of the previous sections and also considering a particular point (x, y), and also being $\widehat{\theta}(x, y)$ as the empirical efficiency obtained for any DEA-type method. For an input-oriented θ, the bias equation could be presented as follows:

$$\text{bias}(\hat{\theta} - \theta) = E(\hat{\theta}(x, y)|(\hat{\theta}x, y) \in \widehat{\Psi}^{\partial}) - [\theta(x, y)|(\theta x, y) \in \Psi^{\partial}] \tag{4.11}$$

In a Monte Carlo procedure $\theta(x, y)|(\theta x, y) \in \Psi^{\partial}$ is known, we could represent it more simply as $\varphi(\theta x, y)$. In an empirical process, the *bootstrap* process is plugged in for the theoretical second term of the right-hand side (not known outside the Monte Carlo simulation world).

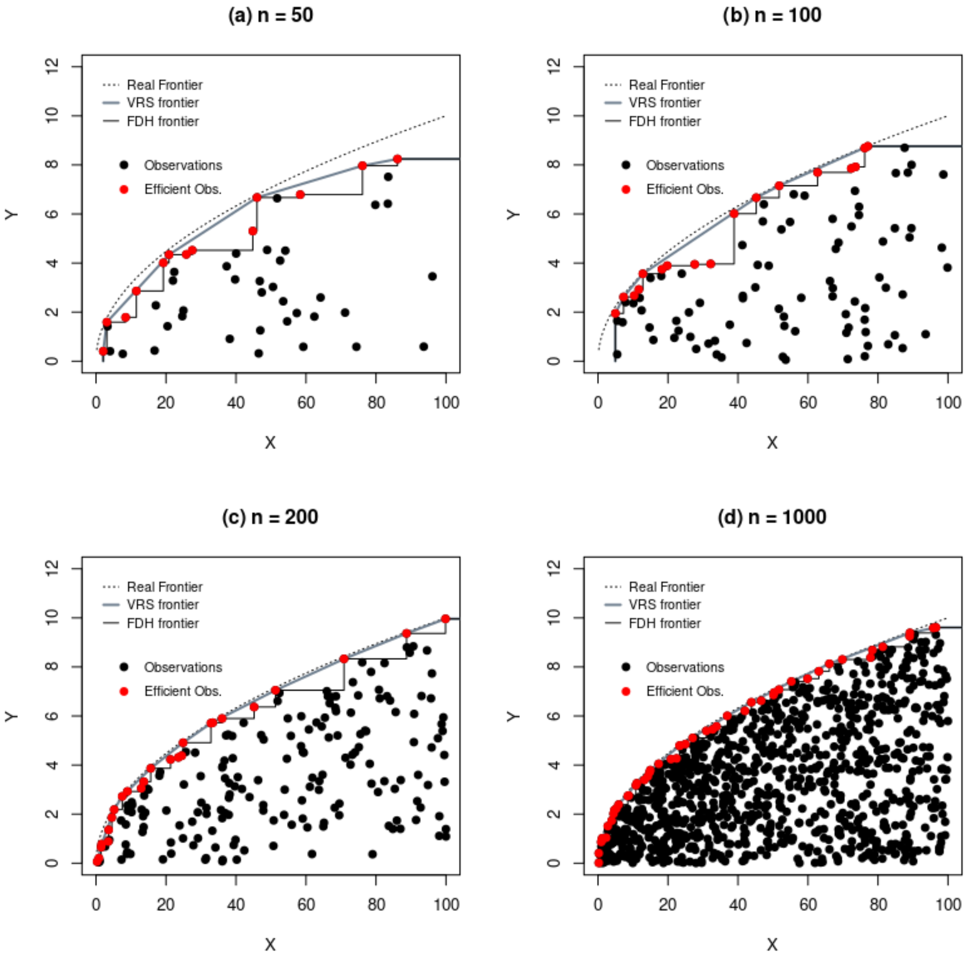


FIGURE 4. Monte Carlo simulations for DGP Model 3, varying n-size (real frontier, VRS and FDH).

We choose three theoretical efficient point to calculate particular bias: $(4, 2)$, efficient in **model 1** and **3** *Data Generating Process*; point $(11.28, 9)$ efficient for **models 2** and **4**; and finally the point $(50, 25)$ efficient for **model 1** but with higher level of input-output.

Table 2 presents the bias by the most well-fitted method (considering returns to scale specification) to each point. As expected, it shows a positive bias for each point. Not so expected was the rapid convergence for those three efficient points, for n from 100 to above the bias is on the second decimal place for the point A, and the third decimal place for points B and C. The variance is getting smaller with slow rate of convergence.

Finally, to make inference, we chose two distribution bias to show the possible interval. Our construction solves some part of the bias inference

TABLE 1. General Monte-Carlo Results to VRS-frontier with four DGP-model specifications (models 1 to 4)

n	Model 1		Model 2		Model 3		Model 4	
	θ_d	$S(\theta_d)$	θ_d	$S(\theta_d)$	θ_d	$S(\theta_d)$	θ_d	$S(\theta_d)$
25	0.599	0.304	0.805	0.173	0.562	0.413	0.889	1.593
50	0.568	0.293	0.745	0.220	0.504	0.592	0.566	0.414
100	0.535	0.294	0.718	0.220	0.401	0.329	0.567	1.261
200	0.671	1.788	0.711	0.215	0.402	0.437	0.412	0.298
500	0.555	1.154	0.716	0.200	0.368	0.301	0.423	0.300
1000	0.524	0.293	0.699	0.208	0.369	0.307	0.419	0.286
2000	0.555	0.274	0.701	0.205	0.352	0.298	0.414	0.292
real θ (target)	0.500		0.667		0.250		0.325	

TABLE 2. Estimating bias for three specific efficient points of interest

n	Point A $(X_i = 4, Y_i = 2)$		Point B $(X_i = 11.28, Y_i = 9)$		Point C $(X_i = 50, Y_i = 25)$	
	θ_d	$S(\theta_d)$	θ_d	$S(\theta_d)$	θ_d	$S(\theta_d)$
25	0.838	7.201	0.035	5.317	0.001	0.027
50	0.396	3.820	0.027	2.738	0.007	0.020
100	0.026	0.502	0.004	1.018	0.010	0.010
200	0.071	0.405	0.002	0.111	0.004	0.005
500	0.030	0.028	0.002	0.031	0.003	0.002
1000	0.004	0.009	0.002	0.000	0.001	0.001
2000	0.001	0.002	0.001	0.000	0.000	0.000
Bias (target)	0.000		0.000		0.000	

5. CONCLUSION

In this article, we reviewed the construction of robust estimators for DEA-type estimators and proposed a method of small displacements of the frontier to avoid bias and obtain more robust estimators. Our approach have some similarities with Bădin and Simar (2009) [23], although it was developed before becoming aware of their work. The presentation focused a lot on the case $(p+q) = 2$ and the Monte Carlo models generated, but a suggested next step soon is to make the developments to more dimensions, following the asymptotic theorems presented in section 3.

The literature on nonparametric frontiers in the last two decades has evolved enormously, in sections 1 and 2 we review much of the pertinent literature. In particular, in comparison with this text is Daouia et al. (2010) [2], which uses techniques more advanced to those proposed here, correcting X through techniques of the Extreme Values Theory (EVT) and

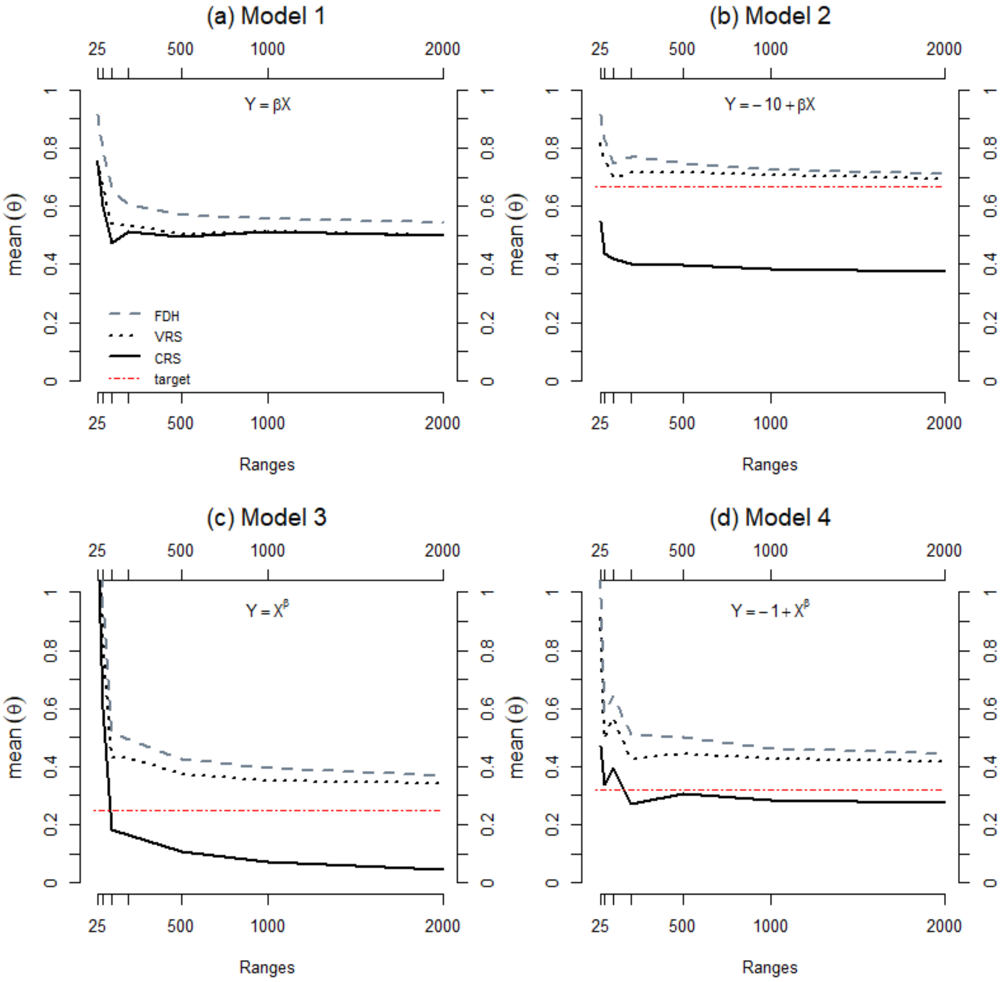


FIGURE 5. Monte Carlo simulations for DGP Model 3, varying n-size (real frontier, VRS and FDH).

focusing on the *Free Disposal Hull* (FDH) estimators. One proposal for future work is to compare our present method with the results of related works.

We may highlight, however, that the technique developed here, based on first and second-order statistics difference, is simpler and more intuitive than some techniques available in the literature [22, 24, 25, 1], but it is also closely related. In fact, there is much to advance in the model proposed here. It is necessary to advance in the suggested *bootstrap* techniques. In addition, more recent articles are dealing more with nonparametric issues for the frontier (as [8]). An idea that we also intend to develop.

There was no space here to deal further with *outliers* issues, such as in Simar (2003) [21] and Sousa and Stošić (2005) [30]. That topic is extremely important for nonparametric boundary estimators. In particular, everything indicates that *outliers* can be very influential in the technique developed here, so this is also a future advance to be pursued.

Regrettably, many questions remain unresolved in this work. In particular, it is necessary to develop the asymptotic properties of the estimator, compare it with more experimentally correlated studies, extend its interpretation to include *bootstrap* methods, incorporate further advances in Extreme Value Theory (EVT), and utilize practical and empirical data, including well-known databases in the field. Addressing these open questions will significantly enhance the applicability of our approach.

We emphasize the importance of correctly specifying the returns to scale specifications. As demonstrated in Figure 5. Poorly specified returns to scale can lead to significant bias, a concept further elaborated in [29]. The method proposed in this study exhibits low bias (compared to DEA models) and reduced variance, as also interpreted according to [29]. Additionally, a notable advantage of this approach is its straightforward extension to larger and applied datasets, as is common in DEA models (usual $(p+q)$ parsimony is maintained). Further research is required to incorporate *environmental* variables and to facilitate the use of these techniques by practitioners.

References

[1] B. U. Park, L. Simar and C. H. Weiner. *The FDH estimator for productivity efficiency scores* Econometric Theory, *16*, 2000, 855–877.

[2] A. Daouia, J. P. Florens and L. Simar. *Frontier estimation and extreme value theory.* Bernoulli, *16*, 2010, 1039–1063.

[3] A. Kneip, L. Simar and P. W. Wilson. *Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models.* Econometric Theory, *24*, 2008, 1663–1697.

[4] R. W. Shephard. *Theory of cost and production functions.* Princeton University Press, 1970.

[5] R. Färe, S. Grosskopf and C. A. K. Lovell. *Production Frontiers.* Cambridge university press, 1994.

[6] C. O'Loughlin, L. Simar and P. W. Wilson. *Methodologies for assessing government efficiency. In Handbook on Public Sector Efficiency.* In Handbook on Public Sector Efficiency (pp. 72-101). Edward Elgar Publishing, 2023.

[7] L. Simar and P. W. Wilson. *Statistical approaches for non-parametric frontier models: a guided tour.* International Statistical Review, *83*, 2015, 77–110.

[8] P. Hall, B. U. Park and S. E. Stern. *On polynomial estimators of frontiers and boundaries.* Journal of Multivariate Analysis, *66*, 1998, 71–98.

[9] D. Deprins, L. Simar and H. Tulkens (1984), Measuring labor-efficiency in post offices, No 571, LIDAM Reprints CORE, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), https://EconPapers.repec.org/RePEc:cor:louvrp:571.

[10] A. Charnes, W. W. Cooper and E. Rhodes. *Measuring the efficiency of decision making units.* European journal of operational research, *2*, 1978, 429–444.

[11] R. D. Banker, A. Charnes and W. W. Cooper. *Some models for estimating technical and scale inefficiencies in data envelopment analysis.* Management science, *30*, 1978, 1078–1092.

[12] A. Moradi-Motlagh and A. Emrouznejad. *The origins and development of statistical approaches in nonparametric frontier models: a survey of the first two decades of scholarly literature (1998–2020).* Annals of Operations Research, *318*, 2022, 713–741.

[13] L. Simar and P. W. Wilson. *Sensitivity analysis of efficiency scores: How to bootstrap in nonparametric frontier models.* Management Science, *44*, 1998, 49–61.

[14] L. Simar and P. W. Wilson. *Estimation and inference in two-stage, semi-parametric models of productive efficiency.* Journal of Econometrics, *136*, 2007, 31–64.

[15] R. C. Sickles and V. Zelenyuk. *Measurement of productivity and efficiency.* Cambridge University Press. 2019.

[16] H. O. Fried, C. A. K. Lovell, S. S. Schmidt and S. Yaisawarng. *Accounting for environmental effects and statistical noise in data envelopment analysis.* Journal of productivity Analysis, *17*, 2002, 157–174.

[17] C. Daraio and L. Simar. *Introducing environmental variables in nonparametric frontier models: A probabilistic approach.* Journal of Productivity Analysis, *24*, 2005, 93–121.

[18] C. Daraio and L. Simar. *How to measure the impact of environmental factors in a nonparametric production model.* European Journal of Operational Research, *223*, 2012, 818–833.

[19] R. Banker, R. Natarajan and D. Zhang. *Two-stage estimation of the impact of contextual variables in stochastic frontier production function models using data envelopment analysis: Second stage OLS versus bootstrap approaches.* European Journal of Operational Research, *278*, 2019, 368–384.

[20] C. Daraio and L. Simar. *Conditional nonparametric frontier models for convex and nonconvex technologies - a unifying approach.* Journal of Productivity Analysis, *28*, 2007, 13–32.

[21] L. Simar. *Detecting outliers in frontier models: A simple approach.* Journal of Productivity Analysis, *20*, 2003, 391–424.

[22] C. Cazals, J. P. Florens and L. Simar. *Nonparametric frontier estimation: a robust approach.* Econometric Theory, *106*, 2002, 1–25.

[23] L. Bădin and L. Simar. *A bias-corrected nonparametric envelopment estimator of frontiers.* Econometric Theory, *25*, 2009, 1289–1318.

[24] Y. Aragon, A. Daouia and C. Thomas-Agnan. *Nonparametric frontier estimation - a conditional quantile-based approach.* Econometric Theory, *21*, 2005, 358–389.

[25] A. Daouia and L. Simar. *Nonparametric efficiency analysis: A multivariate conditional quantile approach.* Journal of Econometrics, *140*, 2007, 375–400.

[26] P. Bogetoft. *Incentives and Productivity Measures* In.: P. Bogetoft. *Non-cooperative planning theory*, chapter 11, pp.247–275 (Vol. 418). Springer-Verlag, 1994.

[27] P. Andersen and N. C. Petersen. *A procedure for ranking efficient units in data envelopment analysis.* Management science, *39*, 1993, 1261-1264.

[28] A. Daouia, J. P. Florens and L. Simar. *Robust frontier estimation from noisy data - A Tikhonov regularization approach.* Econometrics and Statistics, *14*, 2020, 1–23.

[29] A. Kneip, L. Simar and P. W. Wilson. *When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores.* Econometric Theory, *31*, 2015, 394–422.

[30] M. C. S. Sousa and B. Stošić *Technical efficiency of the Brazilian municipalities: correcting nonparametric frontier measurements for outliers.* Journal of Productivity analysis, *24*, 2005, 157–181.

Department of Economics, Programa de Pós-Graduação em Economia Aplicada (PPEA-UFOP), Universidade Federal de Ouro Preto, CEP: 35420-000, Mariana, MG, Brazil

Email address: victor.delgado@ufop.edu.br