Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP07/2019

Data Envelopment Analysis and Business Analytics:

The Big Data Challenges and Some Solutions

Valentin Zelenyuk

Date: May 2019

School of Economics
University of Queensland
St. Lucia, Qld. 4072
Australia

ISSN No. 1932 - 4398

Data Envelopment Analysis and Business Analytics: The Big Data Challenges and Some Solutions

Valentin Zelenyuk*

May 8, 2019

Abstract

The goal of this article is three-fold. The first goal is to present a concise review of Data Envelopment Analysis (DEA) for the more general Business Analytics (BA) community. The second goal of this paper is to discuss the key aspect (and thus the key challenge) of BA—the 'big data'—to the DEA community, which besides a few exceptions, appears to have been largely circumventing this area, despite it gaining more and more attention in other areas of research and practice. The third, and most important, goal of this paper is to discuss possible solutions to the 'big data' problem related to the large dimensions in the context of DEA. To achieve the latter goal, we presented some theoretical grounds and performed a new simulation study to explore the price-based aggregation as a solution to address one of the key challenges of the 'big data' problems for DEA—the immense dimensionality problem.

Keywords: Data Envelopment Analysis; Productivity; Efficiency; Business Analytics; Big Data

*Address: School of Economics and Centre for Efficiency and Productivity Analysis, The University of Queensland, Australia. Email: v.zelenyuk@uq.edu.au. **Acknowledgment**: I thank the Editor and three anonymous referees for many fruitful comments that helped improving this paper substantially. I also thank organizers of APPC2018 and Robin Sickles for their invitation, which inspired me to finish this particular paper I was working on for a long time. I also thank and acknowledge the financial support from the Australian Research Council (from the ARC Future Fellowship grant FT170100401) and from The University of Queensland. I also thank Bao Hoang Nguyen, Duc Manh Pham and Evelyn Smart for proofreading and their feedback. These individuals and organizations are not responsible for the views expressed in this paper.

1 Introduction

Business Analytics (BA) is a catchy and trendy term frequently heard these days in both academia and especially in real business practices. As with some other fields of research, there is no singular unanimously agreed definition of what BA is. The most concise (at least among the widely cited) of the definitions appears to be the one from INFORMS, that defines

> "... analytics as the scientific process of transforming data into insights for the purpose of making better decisions".[1]

Although being quite concise, a natural limitation of this definition is its not being very precise. Indeed, many areas of research use some scientific methods of transforming data for making better decisions, and so would therefore qualify to be labeled as BA according to this definition. An important value of this definition, however, is that it emphasizes the great broadness (and, at least potentially, a great depth) of this area and therefore its great potential, especially because it is supported by the demands of real businesses, striving to improve their performance. Meanwhile, the research and practice of various aspects of performance that from the early stages involved operations research (OR) and management science (MS), economics, statistics and information technologies (IT), as being a symbiosis of these, is referred to as efficiency and productivity analysis (EPA).[2]

One of the main approaches in EPA is *Data Envelopment Analysis* (DEA). This approach was brought to a wider audience by this very journal, by two patriarchs of OR/MS, Abraham Charnes and William Cooper and their student, Edwardo Rhodes, through their seminal paper, Charnes et al. (1978). This work was in turn inspired by the fundamental, yet relatively unknown at that time, work of a prominent economist, Michael J. Farrell (Farrell (1957)) and, at least indirectly, by the works of the Nobel Laureates in Economics, Wassily Leontieff,[3] Leonid Kantorovich,[4] Gerard Debreu[5] and Tjalling Koopmans,[6] as well as by Ronald Shephard,[7] who many think could have been a Nobel Laureate too, if he had lived longer. In turn, some of these works were heavily influenced by ideas from brilliant mathematicians of the last century, such as John von Neumann[8]) and George Dantzig[9] who were among the first (as was Kantorovich) to exploit the great potential of modeling economic phenomena via systems of linear equations or inequalities. It is also worth noting that in the early stages of the literature, this approach was referred to as *Activity Analysis Modeling* (AAM).

Since these early works, many variations of DEA have been proposed and scrutinized in the literature and extensively used by academics, businesses and government researchers to tackle various challenging research questions. These days, DEA is one of the most popular methods for analyzing efficiency and productivity of various systems: firms, banks, hospitals, industries, regions, countries, etc. The four decades after Charnes et al. (1978) have been very fruitful for the DEA method with many theoretical developments and empirical findings in a myriad of papers.[10]

Lustig et al. (2010) distinguished three key categories of BA: (i) descriptive analytics (i.e., 'what happened?'), (ii) predictive analytics (i.e., 'what is likely to happen?') and (iii) prescriptive analytics (i.e., 'what should be

[1] https://www.informs.org/About-INFORMS/News-Room/O.R.-and-Analytics-in-the-News/Best-definition-of-analytics

[2] A recognition that EPA is part of BA appears to be gaining a wide agreement in the research community, e.g., with top field journals preparing special issues on this aspect. We thank anonymous referee for this insights.

[3] E.g., see Leontief (1925).

[4] E.g., see Kantorovich (1960) (the original article of this English version was published in 1939).

[5] E.g., see Debreu (1951).

[6] E.g., see Koopmans (1951a,b)

[7] E.g., see Shephard (1953)

[8] E.g., see von Neumann (1945); the original article of this English version was published in 1938.

[9] E.g., see Dantzig (1949), Dantzig (1955), Dantzig (1963) and Dantzig (1982).

[10] E.g., as of 7 Feb 2019, Google Scholar yielded about 109,000 results when searching for "data envelopment analysis". For extensive reviews of DEA from different perspectives, see Dyson and Thanassoulis (1988), Boussofiane et al. (1991), Allen et al. (1997), Førsund and Sarafoglou (2002), Cook and Seiford (2009) and Emrouznejad et al. (2008) as well as books by Charnes et al. (1994), Ray (2004), Cooper et al. (2007) and Cooper et al. (2011).

done?').[11] An extensive look at the literature on EPA in general, and DEA in particular, suggests that the main focus of research was on (i) and (iii), with great contributions made to these aspects, both from theoretical and from empirical perspectives.[12] Meanwhile, the predictive aspect of it is yet to be realized.[13]

A broad look at the literature also suggests that while naturally fitting into the spectrum of what is now typically referred to as BA (at least in a broad sense), besides a few exceptions, the literature of DEA has been largely circumventing one of the key aspects of modern business analytics—the 'big data'—the aspect that has been reshaping and revolutionizing the real world these days.[14] It is, however, very likely that this particular aspect will substantially influence the future value of DEA, and EPA in general, as well as its use by practitioners and by Artificial Intelligence (AI). Therefore, the main goal of this paper is to explore this important aspect in the context of DEA and suggest some solutions to the challenges it poses. To do so, we will leverage on existing works on aggregation in economics and perform a simulation study, to confirm the optimistic theoretical conclusions that can help in addressing some of the 'big data' challenges posed by the rapidly changing modern era for the DEA researchers.

2 Envelopment Analytics

While economic theorists (e.g., Debreu (1951), Koopmans (1951a,b), Farrell (1957)) provided a philosophical and mathematical foundation for EPA, it is mainly the operations research and management science (OR/MS) literature that empowered this area of research in a practical aspect, as well as advancing its computing aspect, and thus enabled its widespread use not just in academic but also in business and government driven research. Indeed, the now famous work of Farrell (1957) got the attention it deserved only two decades after it was published. This attention came mainly due to the works in OR/MS literature, sparked with the seminal work of Charnes et al. (1978) that generalized Farrell's approach to the multi-output case and connected it to the powerful linear programming theory. The latter work also branded this approach with a catchy name, DEA, and gave it the momentum that has persisted through many other works continuing to the present day.[15]

2.1 Basic DEA models

To be more precise, it is worth recalling that Charnes et al. (1978) started with a fractional programming problem, where the objective function was a type of productivity index (the ratio of aggregate output to aggregate input) for a firm $j \in \{1, ..., n\}$ with an allocation (x^j, y^j), given by

$$E^j_{CCR} = \max_{\substack{v_1,...,v_N \\ u_1,...,u_M}} \left\{ \frac{\sum_{m=1}^M u_m y^j_m}{\sum_{l=1}^N v_l x^j_l} \ : \ \frac{\sum_{m=1}^M u_m y^k_m}{\sum_{l=1}^N v_l x^k_l} \leq 1, \ k = 1, ..., n, \right.$$
$$\left. u_m \geq 0, \ v_l \geq 0, \ l = 1, ..., N; \ m = 1, ..., M \right\} \quad (1)$$

where $x^k = (x^k_1, ..., x^k_N) \in \mathbb{R}^N_+$ and $y^k = (y^k_1, ..., y^k_M) \in \mathbb{R}^M_+$ are observations on inputs and outputs, respectively, of decision making unit (DMU) k and $u' = (u_1, ..., u_M)$ and $v' = (v_1, ..., v_N)$ are optimization variables (sometimes

[11] Also see related discussion in Kohavi et al. (2002); Chen et al. (2012); Mortenson et al. (2015)

[12] E.g., one may notice that most of the papers cited in, for example, such reviews as Cook and Seiford (2009) and Emrouznejad et al. (2008) indeed fall largely into these two categories.

[13] Among a few exceptions are works of Barr et al. (1994), Wheelock and Wilson (1995, 2000), Daskovska et al. (2010) .

[14] E.g., Chen et al. (2012); Marr (2015); Mortenson et al. (2015) for related general discussions on big data, and Barr and Durchholz (1997), Dula (2008), Zhu et al. (2018) and Khezrimotlagh et al. (2019) for the DEA context. Also see Misiunas et al. (2016) for related discussions.

[15] Here, it is also important to acknowledge the contributions from the group of agricultural economics students at the University of California at Berkeley, most prominently Seitz (1970, 1971). We thank the anonymous referee for this insight.

referred to as 'multipliers').[16]

Importantly, Charnes et al. (1978) proved that (1) is equivalent to the problem of estimating the input oriented Debreu-Farrell technical efficiency measure of an activity (x^j, y^j) via a generalized version of the activity analysis model formulated by Farrell (1957), given by

$$\begin{aligned} \widehat{ITE}(x^j, y^j|\text{DEA-CRS}) \quad \equiv \quad & \min_{\theta,\, z^1,\ldots,z^n} \theta \\ \text{s.t.} \quad & \\ & \sum_{k=1}^{n} z^k y_m^k \geq y_m^j,\ m = 1, ..., M, \\ & \sum_{k=1}^{n} z^k x_l^k \leq \theta\, x_l^j,\ l = 1, ..., N, \\ & \theta \geq 0,\ z^k \geq 0,\ k = 1, ..., n. \end{aligned} \tag{2}$$

This particular formulation is more common in economics literature (largely due to its connection to works of Gerard Debreu and Tjalling Koopmans (e.g., Debreu (1951), Koopmans (1951a,b)), and is often referred to as the 'envelopment form of DEA' (here, with assumption of CRS, additivity and free disposability).

Similarly, the DEA-*estimator of the Debreu-Farrell output oriented technical efficiency* score of any activity (x^j, y^j), under the assumptions of CRS (and additivity and free disposability of all outputs and all inputs), is given by

$$\begin{aligned} \widehat{OTE}(x^j, y^j|\text{DEA-CRS}) \quad \equiv \quad & \max_{\theta,\, z^1,\ldots,z^n} \theta \\ \text{s.t.} \quad & \\ & \sum_{k=1}^{n} z^k y_m^k \geq \theta\, y_m^j,\ m = 1, ..., M, \\ & \sum_{k=1}^{n} z^k x_l^k \leq x_l^j,\ l = 1, ..., N, \\ & \theta \geq 0,\ z^k \geq 0,\ k = 1, ..., n. \end{aligned} \tag{3}$$

As is required theoretically for CRS technologies, we have $\widehat{ITE}(x, y) = 1/\widehat{OTE}(x, y)$ for any (x, y). However, for a non-CRS case, the two orientations may yield different results, both quantitatively and qualitatively (i.e., imply different conclusion or policy implication).[17]

The first wave of modifications of the DEA-CRS model focused on relaxing assumptions of CRS and of convexity, and goes back to at least Afriat (1972). It was further formalized and refined by Diewert and Parkan (1983); Färe et al. (1983), Banker et al. (1984), Deprins et al. (1984), Färe and Grosskopf (1985) and Banker and Morey (1986a,b), Petersen (1990); Bogetoft (1996); Bogetoft et al. (2000) to mention just a few.

Out of various proposals, the modifications that remained popular are the DEA-VRS and the DEA-NIRS models, which simply amount to additional constraints $\sum_{k=1}^{n} z^k = 1$ or $\sum_{k=1}^{n} z^k \leq 1$, respectively, to the envelopment form of DEA-CRS.

Furthermore, the so-called Free Disposal Hull (FDH) approach, proposed by Deprins et al. (1984), has also

[16] This formulation of DEA appears much more frequently in the OR/MS literature, and is usually referred to as the 'CCR model' and sometimes called the 'multiplier form of DEA', under assumptions of constant returns to scale (CRS), additivity and strong disposability of inputs and outputs.

[17] E.g., it is possible that for the same data set and same production model, the input (output) orientation suggests full technical efficiency while the output (input) orientation may imply high technical inefficiency. Due to this phenomenon, explanations of reasoning for the particular choice or a robustness check with respect to different orientations are typically advised in the non-CRS cases.

received more attention, especially recently. While not referred to as a DEA estimator, but with its own name, this approach is also an envelopment estimator and can be viewed as a symbiosis of the integer programming and the linear programming problems: It can be formulated as the DEA-VRS, except that the constraints "$z^k \geq 0,\ k = 1, ..., n$" are replaced with "$z^k \in \{0, 1\},\ k = 1, ..., n$". For instance, for an allocation (x^j, y^j), such FDH estimator for the input oriented Debreu-Farrell technical efficiency score is given by

$$\begin{aligned} \widehat{ITE}(x^j, y^j|\text{FDH}) \quad &\equiv \quad \min_{\theta, z^1, ..., z^n} \theta \\ &\text{s.t.} \\ &\sum_{k=1}^{n} z^k y_m^k \geq y_m,\ m = 1, ..., M, \\ &\sum_{k=1}^{n} z^k x_l^k \leq \theta\, x_l,\ l = 1, ..., N, \\ &\sum_{k=1}^{n} z^k = 1, \\ &\theta \geq 0,\ z^k \in \{0, 1\},\ k = 1, ..., n. \end{aligned} \tag{4}$$

while for the output oriented Debreu-Farrell technical efficiency score, it is given by

$$\begin{aligned} \widehat{OTE}(x^j, y^j|\text{FDH}) \quad &\equiv \quad \max_{\theta, z^1, ..., z^n} \theta \\ &\text{s.t.} \\ &\sum_{k=1}^{n} z^k y_m^k \geq \theta\, y_m,\ m = 1, ..., M, \\ &\sum_{k=1}^{n} z^k x_l^k \leq x_l,\ l = 1, ..., N, \\ &\sum_{k=1}^{n} z^k = 1, \\ &\theta \geq 0,\ z^k \in \{0, 1\},\ k = 1, ..., n. \end{aligned} \tag{5}$$

The above formulations are the DEA estimators of the true input and output oriented Debreu-Farrell technical efficiency measures or scores, defined respectively as

$$ITE(x, y|\Psi) \equiv \min_{\theta}\{\theta\ :\ (\theta x, y) \in \Psi\}, \tag{6}$$

and

$$OTE(x, y|\Psi) \equiv \max_{\theta}\{\theta\ :\ (x, \theta y) \in \Psi\}, \tag{7}$$

where Ψ is the technology set, defined in general terms as

$$\Psi \equiv \{(x, y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M\ :\ x \text{ can produce } y\}, \tag{8}$$

which is assumed to satisfy some standard regularity conditions of production theory in economics (e.g., see Shephard (1953); Färe and Primont (1995)).

There are other streams and branches of DEA (and more general envelopment estimators) research, which due

to space limitation we omit here and focus our attention onto these basic models in the context of 'big data'.[18]

2.2 Statistical Underpinnings of DEA: A Path To Accurate BA for Performance Analysis

Many articles describing BA point out that statistical methods and related skills form a critical foundation for successful BA (Kohavi et al. (2002); Lustig et al. (2010); Liberatore and Luo (2011); Chen et al. (2012); Ranyard et al. (2015); Mortenson et al. (2015)). This aspect was also recognized in the DEA literature at a relatively early stage and a lot has been developed since then.

The first breakthrough on statistical properties of DEA was due to Banker (1993) who provided the first proof (for a single-output case in output oriented context, and under certain regularity assumptions) that the DEA approach gives a consistent estimator for the technical efficiency score and pointed out that it belongs to the class of maximum likelihood estimators. This result was soon substantially enriched by Korostelev et al. (1995a,b) who provided a more rigorous proof of consistency of these estimators and demonstrated that they share some optimality properties under fairly general conditions of non-parametric estimation. Importantly, they also derived convergence rates for both DEA and FDH estimators. Specifically, they showed exactly how these rates depend on the dimension of the production model.

The convergence properties and convergence rates for the fully multivariate (multi-input-multi-output) case were first established in the seminal work of Kneip et al. (1998). The most important conclusion of these works is that under certain regularity conditions on the data generating process, the DEA and FDH estimators of technology and related Farrell efficiency measures are consistent, i.e., converge in probability to the truth that they are designed to estimate. However, the speed of convergence of these estimators (at any fixed point (x, y)) depends on the dimension of the production model $(N + M)$ as following

$$\widehat{ITE}(x,y) - ITE(x,y) = O_p(n^{-\kappa}), \text{ and } \widehat{OTE}(x,y) - OTE(x,y) = O_p(n^{-\kappa}) \tag{9}$$

where $O_p(n^{-\kappa})$ denotes the order of magnitude and κ depends on the estimator.[19] Specifically, for DEA-CRS we have $\kappa = 2/(N + M)$, slightly slower for DEA-VRS, being $\kappa = 2/(N + M + 1)$ and even slower for FDH, being $\kappa = 1/(N + M)$. In all these cases, one can see that the higher the dimension, the slower the rate of convergence to the truth. This means that such 'big data' cases where the dimension is very high relative to the number of observations may imply 'big problems' for the DEA and especially for FDH. We will come back to this in the next section.

The first breakthrough on the limiting distribution of the DEA-CRS estimator was made by Gijbels et al. (1999), but only for the 1-input-1-output case. Almost a decade later, Kneip et al. (2008) extended this result to the fully multivariate case for the DEA-VRS estimator and, importantly, also proved consistency of various bootstrap procedures. Soon after this, the asymptotic distribution for the case of the DEA-CRS estimator was established by Park et al. (2010). Meanwhile, the limiting distribution of the FDH estimator for the fully multivariate case was first derived by Park et al. (2000), while consistency of the bootstrap for the FDH estimator was proved by Jeong and Simar (2006).

The most important conclusions of these works on the asymptotic theory of DEA and FDH is four-fold:

(i) under certain regularity assumptions on the data generating process, the DEA and FDH estimators of technology and related Farrell efficiency measures have well-behaved (non-degenerate) limiting distribution,

[18] It is also worth noting that while the original and the most common use of DEA is for estimation of efficiency in production context (as we focus on it here), it also found its uses in other areas of research, e.g., as a tool for benchmarking, for developing composite indexes, etc. (e.g., see Cook et al. (2014, 2017) and Sherman and Zhu (2013)). We thank the anonymous referee for this insight.

[19] To be precise about the meaning of $O_p(n^{-\kappa})$, recall that if $\{\xi_n\}$ is a sequence of random variables and $\{c_n\}$ is a sequence of non-stochastic finite numbers, then $\xi_n = O_p(n^{-\kappa})$ if $(\xi_n/n^{-\kappa}) - c_n \to 0$ in probability, as $n \to \infty$.

(ii) the speed of convergence (in probability and in distribution) depends on the dimension of the production model,

(iii) the characterization of the limiting distribution is too complex and depends on many unknown parameters,

(iv) various bootstrap approaches that are theoretically justified (in terms of statistical properties) are available for estimating and correcting the statistical bias of efficiency estimates, for estimating confidence intervals and for conducting statistical tests.

Some of these works proposed seemed to be the only valid way to perform inference with the DEA and FDH estimates.

Well before the proofs of bootstrap consistency for DEA and FDH were discovered by Jeong and Simar (2006) and Kneip et al. (2008), the ideas of adapting bootstrap (from Efron (1979), Bickel and Freedman (1981); Beran and Ducharme (1991); Hall (1992)) for statistical analysis and inference in DEA and FDH frameworks was first explored by Simar (1992) and further developed by Simar and Wilson (1998, 2000a,b), and more recently studied by Kneip et al. (2008), Simar and Zelenyuk (2006, 2007); Daskovska et al. (2010); Simar et al. (2012); Daraio and Simar (2014), to mention a few.

A particular popular sub-stream here includes works focusing on the so-called 'two-stage DEA analysis', where the first stage involves DEA and the second stage involves their statistical analysis, such as analysis of distributions of efficiency scores (Simar and Zelenyuk (2006)), analysis of aggregate efficiency scores (Simar and Zelenyuk (2007)) and, what seems to be the most popular, analysis of explanatory factors of efficiency scores with a regression. For the latter framework, the approach proposed by Simar and Wilson (2007) appears to be the most popular in practice, which uses maximum likelihood-based truncated regression with double-bootstrap that has the first mode of bootstrap correcting the statistical bias of the DEA estimation, while the second mode of bootstrap estimates the confidence intervals for the regression coefficients. A competing approach here is due to Banker and Natarajan (2008) that relies on a different set of assumptions and involves ordinary least squares-based regression, with standard inference (e.g., see related discussion by Simar and Wilson (2011)).

An important assumption in both approaches is about the specific functional form of the relationship between the true efficiency scores and explanatory factors (typically assumed to be linear in parameters).[20] Also, it is important to note that both approaches require the so-called separability assumption on the data generating process—the condition that requires that the explanatory (sometimes called 'environmental') variables only influence the efficiency and not the frontier. This assumption can be mitigated using alternative approaches, e.g., those based on the concept of partial frontiers and conditional efficiency measures (Cazals et al. (2002), Daraio and Simar (2005, 2007b,a, 2014) and Bădin et al. (2010, 2012)) or using the non-parametric stochastic frontier approach (e.g., Kumbhakar et al. (2007), Park et al. (2015) and Simar et al. (2017)).[21]

More discussion and details can be found in the reviews by Simar and Wilson (2013, 2015). Most recently, Kneip et al. (2015) discovered very important results. They showed that the standard central limit theorems for averages of efficiency scores fail when the true efficiency scores are replaced with their DEA or FDH estimates (due to the statistical bias of DEA and FDH estimates). They also derived new central limit theorems based on jackknife-type bias correction. These results were further extended in the follow up works of Kneip et al. (2016); Daraio et al. (2017) and Simar and Zelenyuk (2018), who also proposed several statistical tests.

Overall, the recent developments regarding the statistical aspect of DEA and FDH have laid a solid foundation for its use in BA, where the statistical aspect is indeed critical.

[20] In principle, one can use a non-parametric version of this method, e.g., adopting the local likelihood approach to truncated regression proposed by Park et al. (2008), though at the cost of the additional 'curse of dimensionality' problem pertinent to nonparametric methods.

[21] Again, as with any kernel-based methods, these approaches are also not immune from the 'curse of dimensionality,' although it can be mitigated to some extent by imposing some structure, e.g., the generalized-additive non-parametric structure.

3 'Big Data' and its Challenges for DEA

While there are different definitions of what is 'big data', depending on the context and perspectives, five key and interrelated 'Vs' are typically distinguished there:[22] (i) Variety, (ii) Volume, (iii) Veracity, (iv) Velocity, (v) Value.

The first 'V' typically refers to the great heterogeneity of data—something researchers had been dreaming about for centuries, and finally it arrived—though typically with the challenge that some or all of the data is so heterogenous that it is hard to structure it into a matrix (a key algebraic object researchers are accustomed to operate with while modeling the phenomena of interest). Even after using various approaches for data preparation in order to overcome the challenge of unstructured data, the resulting matrix of data may often be too large in terms of the dimension of variables, and can be especially challenging if larger than the number of observations (sample size n). This is the main issue we will focus on in this article.

The second 'V' refers to the gigantic amount of information in the available data, which is a natural consequence of the first V. This often leads to challenges of storage and usage of the whole data simultaneously (as has been typically done before).

The third 'V' typically refers to such aspects as validity and accuracy, as well as authenticity and legitimacy (ethical and legal). Indeed, while there is often a great heterogeneity, and massive amount of information, some (or perhaps all) of it may be inaccurate or even invalid or unethically or even illegally sourced, which can put any research under question. For the context of DEA, where researchers used to assume (somewhat naively) that the information on inputs and outputs is accurate, with no statistical noise, it is an especially sensitive matter. A big toolbox of various statistical methods is available and continues to be developed, along with machine learning methods, to check, validate and possibly refine the data and the DEA community must get on top of such methods to address these challenges. Some work has been done in this area, e.g., see the works on stochastic DEA and stochastic FDH (e.g., see Charnes and Neralic (1990); Olesen and Petersen (1995); Cooper et al. (1998); Sengupta (1998); Cooper et al. (2004); Simar and Zelenyuk (2011), to mention a few and the recent review by Olesen and Petersen (2016)).

The fourth 'V' refers to the speeds with which the data is updated, transferred between the storage and the user and the speed of computation with the massive data. It is mainly determined by the first two Vs and by the current knowledge on computing. For DEA, a method that has always been heavily intensive on computing, this aspect is particularly important when one wants to perform statistical inference, since in DEA one typically needs to use bootstrap or a variation of the jackknife method. To address this challenge, the DEA community must continue harnessing the most recent developments in computing and optimization algorithms and machine learning (including various types of deep learning), involving parallel and cloud computing.

Finally, it is the fifth 'V' that is ultimately the most important for the practitioners, and its extent and magnitude depends on the other 'V's, potentially in different ways depending on the context and how the related challenges of each of the 'V's are addressed. In other words, the fifth 'V' can also be viewed as a 'derivative' of the other four 'V's—a positive part of them that must be extracted from those 'V's by properly addressing their challenges. How well the DEA is able to capitalize on this 'V' from different types of 'big data' will also largely determine the value of the DEA method in the future.

Currently, the field of EPA in general, and particularly DEA, is yet to catch up with the pace of expansion of data by taming all the aspects implied by these 'V's—the efforts on this endeavor appear to be quite rudimental at this stage. The area is indeed lagging behind, possibly with the risk of never catching up. To be fair, it is worth noting that similar (or perhaps even bigger) problems are shared by many other sub-fields in OR/MS as is well described and warned by Ranyard et al. (2015) and Mortenson et al. (2015). It appears that even greater lagging behind on harnessing the 'big data' can be observed in many (if not most) sub-fields of economics.[23]

[22] E.g., see Marr (2015) as well as related discussions in Kohavi et al. (2002); Chen et al. (2012); Mortenson et al. (2015)

[23] The CRIW/NBER conferences on 'big data' in 2018 and 2019 are notable exceptions and are examples of a rapid catching up on

The area of statistics, on the other hand, and especially of IT and computation, appear to be, quite naturally, at the forefront of addressing the challenge posed by the 'big data' and, in some sense, they pave the way for the areas of OR/MS and economics to catch up. To the best of our knowledge, out of the relatively few studies on DEA for 'big data' contexts, most of them focus mainly on the case of a large number of DMUs and are done largely from the computational perspective. Some examples of these works include Ali (1993), Barr and Durchholz (1997), Dula (2008), Zhu et al. (2018) and most recently Khezrimotlagh et al. (2019), where more discussion and references can be found on this direction. The main goal of these important studies was essentially to find better (faster and accurate) computational algorithms that can handle data with massive sample sizes (i.e., large numbers of DMUs). By and large, the other aspects of the 'big data' for DEA are yet to receive more attention. An exception is the work of Misiunas et al. (2016), which also focused mainly on the case of a large sample size, yet also discussed the case of relatively large dimensions, and approached it from the perspective of the Artificial Neural Networks (AAN) approach.

Our goal here is to complement the current literature from the economics side of OR, focusing on the problem of large dimensions (inputs and outputs), including the case when this dimension is larger than the number of observations.

3.1 'Curse of dimensionality'

The key reasons for the current lagging behind in exploiting the benefits of the 'big data' for the field are also related to the challenges posed by the other 'V's. Most of these challenges eventually lead to the main challenge of the 'big data'—the incapability of some standard, good-old approaches to handle the massive data and produce useful results. There are many different reasons for such incapability, some of them are purely technical or IT-related (e.g. storage and operating memory of IT and the capacity of the software to operate with large matrices), while others, which we will mainly focus on here, are methodological. The existing methods were originally designed for times when researchers were starving for data, i.e., having a limited number of variables and a small number of DMUs to extract information from.

While there are many methodological issues related to those four 'V's, here we will focus on the problem when the number of explanatory variables (in our case the number of inputs and of outputs) is 'too large' in the sense that it makes the model practically infeasible (e.g., due to computational impossibility) or practically inadequate or simply not very useful (e.g., when the obtained solutions are 'spurious', 'degenerate' or meaningless). In the context of DEA, such a case naturally arises when the number of inputs and outputs ($N + M$) is larger than the number of observations on firms (n), yet it can also arise when $N + M$ are substantially lower than n, even by an order of magnitude, yet still large enough to make each firm unique in its own way and thus each firm may attain a perfect or nearly perfect ($\approx$100%) efficiency score.

In the case of DEA and FDH, this particular problem of large dimensions in the 'big data' may appear quite quickly, even for not so large dimensions. This is especially the case for FDH models a bit less for DEA-VRS models, and a bit less again, yet quite possibly even for moderate dimensions, for the DEA-CRS models. Indeed, the rates of convergence described in (9) clearly show the dependency of the accuracy of estimation on the dimension: the higher the dimension, the slower the rate of convergence to the truth. Recent simulation studies in Kneip et al. (2015), Kneip et al. (2016); Daraio et al. (2017) and Simar and Zelenyuk (2018) clearly suggest that to reach an accurate inference, even for small dimensions like $N + M$ is 3 or 4, one needs quite large sample sizes. What can one say and do when $N + M$ is much larger than that, as is typical in 'big data' contexts?

To get a sense of the problem more formally, recall that according to the statistical convergence rates mentioned in the previous section, under certain conditions, the stochastic difference between the true (or hypothesized)

this aspect in economics in the recent years.

efficiency score and its DEA-CRS estimate is of the order $O_p(n^{-2/(N+M)})$.[24] Compare it to a typical parametric rate $O_p(n^{-1/2})$, sometimes referred to as the 'standard parametric rate'. [25] Now, suppose one wants to analyze a production problem where the number of inputs and outputs is, say 40, then this order becomes $O_p(n^{-1/20})$ for DEA-CRS—extremely slow, in the sense that an extremely large sample size is needed to reach statistically reliable results. More precisely, yet still intuitively, this means that to achieve a similar level of asymptotic accuracy as a method with the standard parametric rate $O_p(n^{-1/2})$ for a sample size of n, the DEA-CRS estimator with a dimension of 40 requires a sample of size of about n^{10}. Furthermore, if the total number of inputs and outputs in the production model is say 200 (as those we consider in the simulations below), reaching a similar level of accuracy as the standard parametric rate would require a sample of size of about n^{50}. To be even more specific, when $n = 1000$, it will require a sample size of 10^{150}—the number much larger than the current scientific estimates of the number of elementary particles in the visible universe! Clearly, some 'meaningful' reduction of the dimension into more manageable numbers of inputs and outputs is greatly needed to make such problems solvable in practice—the topic we discuss next.[26]

3.2 Aggregation of Inputs and Outputs as a Way towards Feasibility

In Section 2 we have outlined the basics of DEA and FDH. It has been implicitly assumed that the data on inputs and outputs, $\{(x^k, y^k)\}_{k=1}^n$, is available for the estimation of technical efficiency. In practice, for most modern production processes, such ideal data availability effectively may create a particular 'big data' problem where the dimension is too large to be practically solvable. Consider for example analyzing the performance of hospitals: many different types of inputs are used to produce hundreds of very different types of healthcare procedures and treatments, which are different types of outputs. Or, consider analyzing the performance of banks: here also many different types of inputs are used to produce many different types of financial products, yielding different returns under different risks.

In general, omitting some of the inputs or some of the outputs when modeling a production process may lead to a misspecification bias and, potentially, misrepresent the true process, possibly yielding misleading conclusions. [27]

On the other hand, even if data on all the inputs and all the outputs can be collected, the curse of dimensionality pertinent to DEA may drastically reduce the discriminatory power of the DEA estimator—e.g., causing many, if not every, observation to obtain the score of unity. So, in this respect, a 'meaningful' aggregation of data is not just useful, but could be the only way to proceed. Indeed, in the empirical literature, one can often see that in the estimation of 'technical' efficiency of economic units via DEA, researchers frequently use *labor cost, material cost, capital cost,* etc., rather than the *physical amounts* of different types of labor (of different quality and at different wages), different types of materials and different types of capital, respectively. A similar example can be made for virtually any other modern industry. In a cross-countries analysis with DEA, the use of gross domestic product as the aggregate output representing a myriad of goods and services produced in a country and the use of aggregate capital, representing a myriad of different types of capital, is another vivid example of circumventing the 'big data' problem related to large dimensions—the problem that was addressed there for many years, although it was not branded as a 'big data' problem. Many applied studies would be just as infeasible in practice without such aggregation.

[24] See Simar and Wilson (2015) for more details.

[25] Note that this rate is also achieved by the DEA-CRS and DEA-VRS estimators for production models where the number of inputs and outputs is 4 and 3, respectively and by an FDH estimator for a one-input-one-output case.

[26] For more of the related discussion, see Wilson (2018). Also see Boussofiane et al. (1991), Dyson et al. (2001) and Ruggiero (2005).

[27] E.g., see related discussion in Boussofiane et al. (1991), Dyson et al. (2001) and Ruggiero (2005).

3.2.1 Index Numbers Approach

Index numbers literature that originated in statistics was the first to tackle the variety of challenges of 'big data' by searching for the best ways of aggregating the data. The origins of this approach go back to at least the 19th century, with the seminal works of Laspeyres (1871), Paasche (1874), and fundamental contributions in the early 20th century by Fisher (1921, 1922), Konüs (1924); Konüs and Byushgens (1926), Frisch (1930, 1936), and Törnqvist (1936).

Most of the early works on index numbers were focusing on the matter of aggregation from a technical or axiomatic perspective—trying to find an index that would satisfy most of the properties. A prominent exception was the work of Konüs (1924); Konüs and Byushgens (1926) who tried to relate the technical indexes advocated by statisticians with the economic theory. These ideas were then expanded on using the fundamental studies on economic index numbers in the 70's, 80's and 90's of the last century, most prominently by Diewert (1971, 1974, 1976, 1978, 1980, 1981, 1983, 1992a,b) and Caves et al. (1982a,b) and Diewert (1992a,b), to mention just a few.[28] These works in turn were inspired by economic theory studies on aggregation by Gorman (1953, 1959), Theil (1954) as well as influenced by more recent works on aggregation in econometrics literature (e.g., see reviews by Stoker (1993) and Blundell and Stoker (2005)).

In their essence, these studies on index numbers were early attempts to deal with the 'big data' problem although this term got coined many years later. In a nutshell, they showed that while the aggregation of data usually comes at the cost of losing some information, some types of aggregations are better than others in various contexts. This is mainly because under certain conditions some of them lose relatively little information or may even not lose any essential information at all, as is the case for the so-called *superlative indexes* and *exact indexes,* respectively.[29]

3.2.2 Correlation-based Approach

This approach is based on the Principal Component Analysis (PCA) or its variations—a generally very useful statistical approach frequently deployed for reducing the dimension of data and widely used in many areas of empirical research.

This approach has been described extensively in many books (e.g., see Härdle and Simar (2015)) and its adaptation to DEA appears to be starting at least from Ueda and Hoshiai (1997), Adler and Golany (2001) and more recently refined in Daraio and Simar (2007a) and Adler and Yazhemsky (2010). This approach was recently revisited and explored further via simulations by Wilson (2018), who also proposed three types of diagnostics for checking when PCA-based dimension reduction may lead to a more accurate efficiency estimation in the DEA and FDH contexts. More details on this approach can be found, for example in Wilson (2018), while we will focus here on the price-based aggregation in DEA and FDH.

3.2.3 Price-based Aggregation Approach for DEA

The impact of aggregation of data on inputs or outputs using prices has also been studied for the context of DEA. Although such studies were not cast explicitly as referring to 'big data', some ideas and conclusions can be generalized to this case as well.

This approach can be viewed from the economic side of OR and in the spirit of aggregation approaches that were done in economic theory for decades, which usually rely (directly or indirectly) on duality theory in economics.

[28] See Diewert (1981), Diewert and Nakamura (1993) and Balk (1998, 2004); Balk and Diewert (2001); Balk (2008) for more details and discussion.

[29] To be more precise, when under some regularity conditions (on technology, preferences, etc.) and for some functional form for the aggregating function, if an index number formula coincides with a theoretical index (e.g., productivity index, cost of living index, etc.) then it is called an *exact index*. In addition, the exact index is called a *superlative index* if the aggregating function is a flexible functional form, in the sense of being at least the second order approximation for the unknown true functions that characterize the phenomenon of interest (e.g., technology, preferences.). E.g., see Diewert and Nakamura (1993) and Balk (2008) and references therein.

It is also in the spirit of the index numbers approach discussed above, especially its economic connotation. We will focus on this approach in this paper.

The key theoretical work here is due to Färe and Grosskopf (1985) that showed the equivalence of DEA solutions for cost efficiency and for technical efficiency when all the inputs were aggregated into total costs, using the same prices.

Another key work on this topic was due to Tauer (2001), who using extensive simulations provided evidence that the aggregation of inputs (and similar is true for outputs) may cause aggregation bias in the Debreu-Farrell type measurement of technical efficiency.[30] This means that the estimates of efficiency scores obtained from disaggregated data may tell a different story from the efficiency estimates for the same firms when the data is aggregated over inputs. For instance, the results of the estimation when one uses the total labor costs might be different from the estimation results where one uses information on the hours worked by each type of labor.

One natural question here would be: Can this aggregation bias be avoided? If yes, then under what conditions? Färe and Zelenyuk (2002) explored this question, deriving necessary and sufficient conditions under which the linear aggregator (the one that is most commonly used in practice) of inputs yields an *unbiased* measurement of efficiency. These conditions, *inter alia*, require no allocative inefficiency and so can be viewed as very strict and hardly met in practice. Another set of important questions arises then: How large is the aggregation bias? Can we hope it is sufficiently small? Under what conditions? And, what is the *direction* of the bias, its size and is there a *bound* for it?

From theoretical perspectives, these questions were further explored by Färe et al. (2004), who generalized some important results from Färe and Grosskopf (1985). We will discuss these results in the next sub-section. The main bottom line from these studies is that the benefit of data aggregation does not come for free in DEA—it usually leads to some aggregation bias, yet they also derived some optimistic results, deriving the bounds of the aggregation bias, which turned out to have some economic meaning and an expectation that such an aggregation bias, on average, is expected to be small and can be justified for practical reasons.

3.3 Cost-Output Model: Impact of Aggregation of Inputs into Total Costs

Let us start by looking at the aggregation issue for the case when *all* inputs are aggregated into *one* variable—the total cost. Specifically, suppose for each firm $k \in \{1, ..., n\}$ in a sample, a researcher observes all outputs that each DMU has produced, $y^k = (y_1^k, ..., y_M^k) \in \mathbb{R}_+^M$. However, instead of observing all inputs that these firms used, $x^k = (x_1^k, ..., x_N^k) \in \mathbb{R}_+^N$, the researcher only has (or chooses to use) information on the total cost (not necessarily the minimal cost) incurred, which is given by

$$c^k = \sum_{l=1}^{N} w_l^k x_l^k, \quad k = 1, ..., n, \tag{10}$$

where $w^k = (w_1^k, ..., w_N^k) \in \mathbb{R}_{++}^N$ are the input prices.

A natural starting point is to assume that all firms face the same exogenous input prices, i.e., $w^k = w$, $\forall k$, which is consistent with the competitive equilibrium theory in economics as well as with the uniform regulatory pricing policy. This assumption is often referred to as the 'Law of One Price' assumption, which of course rarely holds in practice (as many other simplifying assumptions): Even if there is some unique theoretical price for all firms, the actual prices that different firms face are often subject to some deviation. Nevertheless, this assumption is critical for the exact theoretical results presented below. It is also valuable to explore how sensitive the results are to small changes in this assumption, which we will explore through Monte Carlo simulations.

The key questions are 'What if we use the observed total cost, $c^k = \sum_{l=1}^{N} w_l x_l^k$, as a 'proxy' for *all* inputs

[30] Also see related work of Thomas and Tauer (1994).

in a DEA model—what kind of efficiency estimates do we get? Is there any relationship to the estimates in the case when all inputs are observed and used in the estimation?' These questions were answered in the important although largely overlooked work of Färe and Grosskopf (1985), which we follow in this sub-section, focusing on the DEA-VRS model where all inputs and outputs are freely disposable. (Similar results can be obtained for other types of returns to scale as well as for weakly disposable inputs/outputs.) In particular, they showed that the analogue of (2) with VRS assumption, but where the observed total cost is used as a proxy for all inputs, i.e.,

$$\begin{aligned}
\widehat{ITE}_c(y^j, c^j) \quad \equiv \quad & \min_{\theta,\, z^1,\ldots,z^n} \theta \\
\text{s.t.} \quad & \\
& \sum_{k=1}^{n} z^k y_m^k \;\geq\; y_m^j,\; m = 1, ..., M, \\
& \sum_{k=1}^{n} z^k c^k \;\leq\; \theta\, c^j, \\
& \sum_{k=1}^{n} z^k = \; 1, \\
& \theta \geq 0,\; z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{11}$$

is equivalent to the following optimization problem

$$\begin{aligned}
\widehat{ITE}_c(y^j, c^j) \quad \equiv \quad & \min_{z^1,\ldots,z^n} \frac{1}{c^j} \sum_{k=1}^{n} z^k c^k. \\
\text{s.t.} \quad & \\
& \sum_{k=1}^{n} z^k y_m^k \;\geq\; y_m^j,\; m = 1, ..., M, \\
& \sum_{k=1}^{n} z^k = \; 1, \\
& z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{12}$$

Meanwhile, recall the definition of the *(minimal) cost function*, given by

$$C(y, w|\Psi) \equiv \min_x \{wx \;:\; (x, y) \in \Psi\}. \tag{13}$$

The DEA formulation of this function for a particular observation (y^j, w) is given by

$$\begin{aligned}
\widehat{C}(y^j, w) \quad \equiv \quad & \min_{\substack{x_1,\ldots,x_N \\ z^1,\ldots,z^n}} \sum_{l=1}^{N} w_l x_l, \\
\text{s.t.} \quad & \\
& \sum_{k=1}^{n} z^k y_m^k \;\geq y_m^j,\; m = 1, ..., M, \\
& \sum_{k=1}^{n} z^k x_l^k \;\leq x_l,\; l = 1, ..., N, \\
& \sum_{k=1}^{n} z^k = 1, \\
& z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{14}$$

and is equivalent to the following optimization problem

$$\begin{aligned} \widehat{C}(y^j, w) \quad &\equiv \quad \min_{z^1,\ldots,z^n} \sum_{k=1}^{n} z^k c^k \\ \text{s.t.} \quad & \\ & \sum_{k=1}^{n} z^k y_m^k \;\geq\; y_m^j, \;\; m = 1, ..., M, \\ & \sum_{k=1}^{n} z^k = \; 1, \\ & z^k \geq 0, \;\; k = 1, ..., n. \end{aligned} \tag{15}$$

Comparing (15) with (12), they arrive at the following important result

$$\widehat{CE}(y^j, w, x^j) \equiv \frac{\widehat{C}(y^j, w)}{c^j} = \widehat{ITE}_c(y^j, c^j), \;\; j = 1, ..., n. \tag{16}$$

That is, when all firms face the same input prices and the same technology, the DEA estimate of the 'technical' efficiency score obtained from the data where all inputs are 'proxied' by the observed total costs, $\widehat{ITE}_c(y^j, c^j)$, would be equivalent to the DEA estimate of the cost or overall input efficiency $\widehat{CE}(y^j, w, x^j)$.

The main importance of this result is that it opens the door for many applications that might have been infeasible due to the absence of input information or when the dimensionality of the input space is too large in comparison with the available sample size. With the result (16), a researcher can substitute all the unobserved inputs by the observed total cost measure, use the standard DEA estimator, and obtain the estimate of *overall* input efficiency. The resulting estimates, in fact, might even have more valuable information about efficiency from an economic point of view, since they incorporate such economically important information such as (input) prices, the corresponding allocation of inputs and the underlying behavior of the DMUs.

3.4 Revenue-Input Model: Impact of Aggregation of Outputs into Total Revenue

Now we look at the aggregation of *all* outputs into *one* variable—total revenue. Assume for each firm $k \in \{1, ..., n\}$ in a sample, a researcher observes all inputs that each DMU has produced, $x^k = (x_1^k, ..., x_N^k) \in \mathbb{R}_+^N$ but instead of all outputs that these firms produced, $y^k = (y_1^k, ..., y_M^k) \in \mathbb{R}_+^M$, the researcher only has (or chooses to use) information on the total revenue (not necessarily the maximal revenue), which is given by

$$r^k = \sum_{m=1}^{M} p_m^k y_m^k, \quad k = 1, ..., n, \tag{17}$$

where $p^k = (p_1^k, ..., p_N^k) \in \mathbb{R}_{++}^M$ are the output prices.

Again, a natural starting point is to assume that all firms face the same exogenous output prices, which is consistent with the competitive equilibrium theory in economics as well as with the uniform regulatory pricing policy. The key questions here are: 'What if we use the observed total revenue, $r^k = \sum_{m=1}^{M} p_m y_m^k$, as a 'proxy' for *all* outputs in a DEA model—what kind of efficiency estimates do we get? Is there any relationship to the estimates in the case when all outputs are observed and used in the estimation?' Using similar logic as in Färe and Grosskopf (1985), one can derive analogous results for the case of using total revenue as a proxy for using all the outputs. To do so, let us multiply each output inequality constraint m in (3) (with VRS imposed) by the corresponding output price p_m, $(m = 1, \ldots, M)$ and sum these constraints over m, while keeping the other constraints the same, to get

the analogue of $\widehat{OTE}(x^j, y^j)$, but where observed total revenue is used as a proxy for all outputs, i.e.,

$$\begin{aligned}
\widehat{OTE}_r(x^j, r^j) \quad &\equiv \quad \max_{\theta,\, z^1,\ldots,z^n} \theta \\
&\text{s.t.} \\
&\qquad \sum_{k=1}^{n} z^k r^k \;\geq \theta\, r^j, \\
&\qquad \sum_{k=1}^{n} z^k x_l^k \;\leq x_l^j,\; l = 1, ..., N, \\
&\qquad \sum_{k=1}^{n} z^k = 1, \\
&\qquad \theta \geq 0,\; z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{18}$$

Since θ is a scalar, we can rearrange (18), and rewrite it into the following optimization problem

$$\begin{aligned}
\widehat{OTE}_r(x^j, r^j) \quad &\equiv \quad \max_{z^1,\ldots,z^n} \frac{1}{r^j} \sum_{k=1}^{n} z^k r^k. \\
&\text{s.t.} \\
&\qquad \sum_{k=1}^{n} z^k x_l^k \;\leq x_l^j,\; l = 1, ..., N, \\
&\qquad \sum_{k=1}^{n} z^k = \;1, \\
&\qquad z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{19}$$

Now, recall the definition of the *(maximal) revenue function*, given by

$$R(x, p|\Psi) \equiv \max_{y} \{py \;:\; (x, y) \in \Psi\}. \tag{20}$$

The DEA formulation of this function for a particular observation (x^j, p) is given by

$$\begin{aligned}
\widehat{R}(x^j, p) \quad &\equiv \quad \max_{\substack{y_1,\ldots,y_M \\ z^1,\ldots,z^n}} \sum_{m=1}^{M} p_m y_m, \\
&\text{s.t.} \\
&\qquad \sum_{k=1}^{n} z^k y_m^k \;\geq y_m^j,\; m = 1, ..., M, \\
&\qquad \sum_{k=1}^{n} z^k x_l^k \;\leq x_l,\; l = 1, ..., N, \\
&\qquad \sum_{k=1}^{n} z^k = 1, \\
&\qquad z^k \geq 0,\; k = 1, ..., n.
\end{aligned} \tag{21}$$

One can multiply each m^{th} output inequality constraint (21) by the corresponding input price p_m, $(m = 1, \ldots, M)$ and, since we are maximizing $\sum_{m=1}^{M} p_m y_m$, one can sum these inequality constraints over m, while

keeping the other constraints the same, to get an equivalent to an optimization problem given by

$$\begin{aligned} \widehat{R}(x^j,p) \quad &\equiv \quad \max_{z^1,...,z^n} \sum_{k=1}^{n} z^k r^k \\ \text{s.t.} \quad & \\ & \sum_{k=1}^{n} z^k x_l^k \;\leq\; x_l^j,\; l=1,...,N, \\ & \sum_{k=1}^{n} z^k = \; 1, \\ & z^k \geq 0,\; k=1,...,n. \end{aligned} \tag{22}$$

Now, comparing (19) with (22), we arrive at the following conclusion:

$$\widehat{RE}(x^j,y^j,p) \equiv \frac{\widehat{R}(x^j,p)}{r^j} = \widehat{OTE}_r(x^j,r^j),\; j=1,...,n. \tag{23}$$

That is, when all firms face the same output prices and the same technology, the DEA estimate of the 'technical' efficiency score (19) obtained from the data where all outputs are 'proxied' by the observed total revenue would be equivalent to the DEA estimate of the revenue or overall output oriented efficiency obtained from (22) after normalizing it by the observed total revenue of the observation j (i.e., r^j).

Similarly as discussed in the previous section, this result opens the door for many applications that might have been infeasible due to the absence of information on all the outputs or when the dimensionality of the output space is too large in comparison with the available sample size. With the result (23), a researcher can substitute all the unobserved outputs by the observed total revenue measure, use the standard DEA estimator, and obtain the estimate of *overall* output efficiency. The resulting estimates, in fact, might even have more valuable information about efficiency from an economic point of view, since they incorporate such economically important information as output prices and the corresponding allocation of outputs.

3.5 Directions of the Aggregation Bias and its Bound

An immediate implication from the previous subsections is that the aggregation of inputs into total costs or outputs into total revenue yields aggregation bias, since for all $j=1,...,n$, we have

$$\widehat{ITE}_c(y^j,c^j) = \widehat{CE}(y^j,w,x^j) \leq \widehat{ITE}(x^j,y^j) \tag{24}$$

and

$$\widehat{OTE}_r(x^j,r^j) = \widehat{RE}(x^j,p,y^j) \geq \widehat{OTE}(x^j,y^j) \tag{25}$$

Expressions (24) and (25) assert that the aggregation bias is always one way—downward in the input oriented framework and upward in the output oriented framework, i.e., in both cases it will give an overestimation of inefficiency, and this aggregation bias is equal precisely to the estimate of the *allocative* efficiency measure. In some sense, such measures of efficiency might be viewed as superior to the technical efficiency measures, because they contain more information and possess greater discriminative power in distinguishing performance across firms.

If not all inputs are aggregated, but only some, say, all types of labor are aggregated into 'labor cost', and/or all types of capital are aggregated into 'capital cost', then the aggregation bias is also inevitable, except for very peculiar technologies and efficiency measures.[31] A natural question here is: 'How large is the resulting aggregation bias and

[31] This question was explored theoretically by Färe and Zelenyuk (2002), who derived the necessary and sufficient conditions under which the linear aggregation yields an *unbiased* measurement of efficiency.

what is its direction?' It turns out that the aggregation bias is also always downward and is bounded—again by the allocative efficiency. This result has been derived by Färe et al. (2004). To summarize their important results, let us first consider the consequences of aggregation of some $\hat{N} \leq N$ inputs using the linear aggregating function with (exogenous) input prices as weights, i.e., when instead of $x_1^k, ..., x_{\hat{N}}^k$, the researcher is using

$$c_{\hat{N}}^k = \sum_{l=1}^{\hat{N}} w_l x_l^k, \ k = 1, ..., n \tag{26}$$

then, the DEA estimate of the input oriented Debreu-Farrell efficiency measure would be

$$\begin{aligned} \widehat{ITE}_{cx}(y^j, c_{\hat{N}}^j, x_{\hat{N}+1}^j, ..., x_N^j) \quad = \quad & \min_{\theta,\, z^1,...,z^n} \theta \\ \text{s.t.} \quad & \\ & \sum_{k=1}^{n} z^k y_{km} \geq y_m^j, \ m = 1, ..., M, \\ & \sum_{k=1}^{n} z^k x_l^k \leq \theta\, x_l^j, \ l = \hat{N}+1, ..., N, \\ & \sum_{k=1}^{n} z^k c_{\hat{N}}^k \leq \theta\, c_{\hat{N}}^j, \\ & \sum_{k=1}^{n} z^k = 1, \\ & \theta \geq 0, \ z^k \geq 0, \ k = 1, ..., n. \end{aligned} \tag{27}$$

If $\hat{N} = N$, then we are in the case discussed above. If $\hat{N} < N$, then for every observation $j \in \{1, ..., n\}$ we will have

$$\widehat{CE}(y^j, w, x^j) = \widehat{ITE}_c(y^j, c^j) \leq \widehat{ITE}_{cx}(y^j, c_{\hat{N}}^j, x_{\hat{N}+1}^j, ..., x_N^j) \leq \widehat{ITE}(y^j, x^j). \tag{28}$$

Intuitively, if some of the inputs are aggregated as in (26) while others enter the estimation in disaggregated form, then the resulting efficiency measure in (27) is *biased downward* relative to the efficiency measure $\widehat{ITE}(y^j, x^j)$. Moreover, note that the aggregation bias is *bounded* by the size of the allocative (in)efficiency, because $\widehat{CE}(y^j, w, x^j) \leq \widehat{ITE}_{cx}(y^j, c_{\hat{N}}^j, x_{\hat{N}+1}^j, ..., x_N^j)$. Importantly, also note that this aggregation bias is in the opposite direction from the statistical bias and so it has a potential to partially cancel out with it.

By a similar argument it also follows that the more inputs that are aggregated into the sub-cost(s) the greater may be the aggregation bias. Moreover, Färe et al. (2004) showed that the size of the aggregation bias will always be limited by the size of the *allocative* (in)efficiency, whatever the level of aggregation is, because

$$\begin{aligned} \widehat{ITE}_c(y^j, c^j) &\leq \widehat{ITE}_{cx}(y^j, c_{\hat{N}+d}^j, x_{\hat{N}+d+1}^j, ..., x_N^j) \\ &\leq \widehat{ITE}_{cx}(y^j, c_{\hat{N}}^j, x_{\hat{N}+1}^j, ..., x_N^j) \leq \widehat{ITE}(y^j, x^j). \end{aligned} \tag{29}$$

for an integer $d \in \{1, ..., N - \hat{N} - 1\}$.

They also generalized this statement to the case when there is more than one input aggregation into different sub-costs (e.g., labor costs, material costs, etc.), by showing that

$$\begin{aligned}\widehat{ITE}_c(y^j,c^j) &\leq \widehat{ITE}_{cx}(y^j,c^j_{\hat{N}_1},...,c^j_{\hat{N}_h},x^j_{\hat{N}_h+1},...,x^j_N)\\ &\leq \widehat{ITE}_{cx}(y^j,c^j_{\hat{N}_1},...,c^j_{\hat{N}_{h-1}},x^j_{\hat{N}_{h-1}+1},...,x^j_{\hat{N}_h},x^j_{\hat{N}_h+1},...,x^j_N) \qquad (30)\\ &\leq \widehat{ITE}(y^j,x^j)\end{aligned}$$

where $c^k_{\hat{N}_h}=\sum_{l=\hat{N}_{h-1}+1}^{\hat{N}_{h+1}-1} w_l x^k_l,\ k=1,...,n$ for an integer $h\geq 2$ that depends on the level of partitioning of x.

From these aggregation results one can deduce an important implication for an applied work, including the cases with big data: the efficiency estimates from models with aggregated inputs and outputs are 'squeezed-in' between the DEA estimate from the original Farrell (1957) technical efficiency measure and the DEA estimate of the cost efficiency from the other side. If we know that the difference between the two is small—i.e., the allocative inefficiency is small—then the bias due to input aggregation will also be small.

The argument we outlined above can also be extended to the case when some of the outputs are also aggregated into sub-revenues, in addition or without aggregation of some of the inputs. An example would be macroeconomic applications of a Farrell-type measurement—where instead of using all outputs (which is practically infeasible) the aggregated output, such as GDP or GNP, is used instead (e.g., see Färe et al. (1994), Kumar and Russell (2002), etc.). To formalize this argument, consider a case when only some ($\hat{M}\leq M$) outputs are linearly aggregated as

$$r^k_{\hat{M}}=\sum_{m=1}^{\hat{M}} p_m y^k_m,\ k=1,...,n,\ \text{and}\ \hat{M}\leq M. \qquad (31)$$

The result of this aggregation, $r^k_{\hat{M}}$, represents the k^{th} firm's observed sub-revenue for the outputs $(y_1,...,y_{\hat{M}})$. The linear programming formulation to compute the j^{th} firm's efficiency score from the Debreu-Farrell measure is now written as

$$\begin{aligned}\widehat{ITE}_{rcxy}(r^j_{\hat{M}},y^j_{\hat{M}+1},...,y^j_M,c^j_{\hat{N}},x^j_{\hat{N}+1},...,x^j_N) &= \min_{\theta,\,z^1,...,z^n}\theta \qquad (32)\\ &\text{s.t.}\\ &\sum_{k=1}^n z^k y^k_m\geq y^j_m,\ m=\hat{M}+1,...,M,\\ &\sum_{k=1}^n z^k r^k_{\hat{M}}\geq r^j_{\hat{M}},\\ &\sum_{k=1}^n z^k x^k_l\leq \theta\, x^j_l,\ l=\hat{N}+1,...,N,\\ &\sum_{k=1}^n z^k c^k_{\hat{N}}\leq \theta\, c^j_{\hat{N}}\\ &\sum_{k=1}^n z^k=1\\ &\theta\geq 0,\ z^k\geq 0,\ k=1,...,n.\end{aligned}$$

Now, using the same line of reasoning as above, we get

$$\begin{aligned}\widehat{ITE}_{rcxy}(r^j_{\hat{M}+r}, y^j_{\hat{M}+r+1}, ..., y^j_M, c^j_{\hat{N}}, x^j_{\hat{N}+1}, ..., x^j_N) &\leq \widehat{ITE}_{rcxy}(r^j_{\hat{M}}, y^j_{\hat{M}+1}, ..., y^j_M, c^j_{\hat{N}}, x^j_{\hat{N}+1}, ..., x^j_N) \\ &\leq \widehat{ITE}_{cx}(y^j, c^j_{\hat{N}}, x^j_{\hat{N}+1}, ..., x^j_N) \\ &\leq \widehat{ITE}(y^j, x^j)\end{aligned} \tag{33}$$

i.e., the aggregation of outputs may increase (and never decrease) the downward aggregation bias of the efficiency score; moreover, it does so monotonically, i.e., the more outputs are aggregated the larger the aggregation bias might be.

Another similar argument can be stated for the case when there is more than one output aggregation into different sub-revenue. This extension of (33), is formally stated as

$$\begin{aligned}&\widehat{ITE}_{rcxy}(r^j_{\hat{M}_1}, ..., r^j_{\hat{M}_s}, y^j_{\hat{M}_s+1}, ..., y^j_M,\ c^j_{\hat{N}_1}, ..., c^j_{\hat{N}_h}, x^j_{\hat{N}_h+1}, ..., x^j_N) \\ &\leq \widehat{ITE}_{rcxy}(r^j_{\hat{M}_1}, ..., r^j_{\hat{M}_{s-1}},\ y^j_{\hat{M}_{s-1}+1}, ..., y^j_M, c^j_{\hat{N}_1}, ..., c^j_{\hat{N}_h}, x^j_{\hat{N}_h+1}, ..., x^j_N)\end{aligned} \tag{34}$$

where $r^j_{\hat{M}_s} = \sum_{m=\hat{M}_{s-1}+1}^{\hat{M}_{s+1}-1} p_m y^k_m,\ k = 1, ..., n$ for an integer $s \geq 2$ that depends on the level of partitioning of y.

Moreover, in the case when all outputs are aggregated, i.e., when $\hat{M} = M$, the lowest bound would be $\widehat{ITE}_{rc}(r^j, c^j)$. Notably, under the CRS assumption (which is a common assumption especially in growth measurement) this bound has a simple interpretation—it is the j^{th} firm's observed revenue-cost ratio, normalized by the highest revenue-cost ratio in the sample, which we denote by $r^{*k'}/c^{*k'}$, i.e.,

$$\widehat{ITE}_{rc}(r^j, c^j|CRS) = \left(\frac{r^j}{c^j}\right) \widehat{ITE}_{rc}(1, 1|CRS) = \left(\frac{r^j}{c^j}\right) / \left(\frac{r^{*k'}}{c^{*k'}}\right). \tag{35}$$

Thus, in this limiting case, we get the so-called "return to the dollar" measure of performance advocated by Georgescu-Roegen (1951), normalized by the highest "return to the dollar" in the sample. Moreover, note the resemblance of (35) to the CCR efficiency measure, which is also a (reciprocal) ratio of aggregate output to aggregate input, except that it allows shadow prices to optimize according to the data on other observations, while this one restricts the prices to be the observed ones, and common across all firms.

Importantly, note that this ratio bounds from below the efficiency score $\widehat{ITE}_{rc}(r^j, c^j|VRS)$ computed under variable returns to scale and bounds from below all the other inequalities for the DEA-estimated input oriented technical efficiency scores.

It is worth noting that although the aggregation bias for every firm is bounded in the same fashion, its size may (and most likely will) vary from firm to firm—depending on the firm's allocative inefficiency, implying that the *ranking* of firms by their efficiency scores may be different at different levels of aggregation.[32] Thus, empirical researchers using DEA on aggregated inputs or outputs should hope not only to obtain the same estimates as if they were using disaggregated data, but even to obtain the same efficiency-based ranking of firms. This however does not mean that one should not do an empirical work on aggregated data—because often there is no other data! Or, even if there is detailed data, the curse of dimensionality may prevent the researcher from using it, forcing him/her to use aggregate data. What is meant however, is that the interpretation should not misleadingly be referred to as the 'pure' technical efficiency measure or to the estimate of the 'pure' technology set. Instead, researchers may want to explicitly emphasize that some of the inputs or outputs are replaced with their respective (sub)cost and

[32] Thomas and Tauer (1994) confirmed this phenomenon with real data on dairy farms and Tauer (2001) explored this question further via an extensive simulation-based study.

(sub)revenue proxies, which in fact might even be desirable. This is because the revenue and costs incorporate such important economic information as prices.

Such aggregated measurement of inputs and outputs may give some justification for pooling the data over, say, different industries or countries to estimate one unique best practice frontier. In this sense, technology set characterization, Ψ, would not bear the engineering, but a broader meaning—this would be a set characterizing possibilities to make (sub)revenues out of investments into major inputs, or in colloquial terms, to 'use cash to make more cash', regardless of the particular engineering, managerial and other features of the businesses. And, perhaps most importantly, such pooling of data over industries often can give more confidence in the precision of the DEA estimation—due to the substantial increase in the sample size—which is very important for obtaining accurate estimates. Later, upon obtaining the efficiency estimates, one could analyze their differences in means, variances or entire distributions across these pooled industries.

4 A Simulation Study[33]

Here, our goal is to explore the sensitivity of results using three main estimators, DEA-CRS, DEA-VRS and FDH, to various levels of aggregation of inputs and outputs. The theory outlined above clarifies that there is a bias due to aggregation, the source of which is derived from some or all of allocative efficiency and it is bounded by it. An important question for practice is how big is this aggregation bias or how much information is lost due to such aggregation. Of course, a precise answer would differ across different data and different contexts, yet some useful insights can be obtained from simulations or Monte Carlo experiments and so here we present typical results from various scenarios of fairly extensive simulations we performed.

4.1 A Data Generating Process

In all the experiments we will consider the production technology characterization of the following form

$$\Psi \quad = \quad \left\{ (x,y) \ : \ \left(\sum_{m=1}^{M} \beta_m (y_m)^2 \right)^{1/2} \leq \left(\prod_{l=1}^{N} (x_l)^{\alpha_l} \right)^{\gamma} \right\} \tag{36}$$

where $\alpha_l \geq 0$ and $\beta_m \geq 0$ and we set $\sum_{l=1}^{N} \alpha_l = 1$ to ensure linear homogeneity of the distance functions, yet allow for $\gamma \in (0,1]$ as a parameter of scenario choice for investigating different degrees of the returns to scale. The r.h.s. of the inequality in (36) is the well-known Cobb-Douglas functional form, which is very popular in theory and practice and is often used in simulation studies. It can be understood as the first order approximation (in logs) of the true production function. The fact that the formulation is multiplicative (or linear in logs) allows for some curvature in the original units of measurement and models both substitutability and complementarity of inputs, unlike such extremes as linear or Leontieff-type technologies. For the output-side of the technology characterization we use the weighted Euclidian norm where, without loss of generality, we normalize $\sum_{m=1}^{M} \beta_m = 1$. One could also try the Cobb-Douglas form for the output side here, yet this will make the output sets non-convex and also will not allow zero outputs (thus violating standard production axiom that inactivity is possible).

In principle, one could use higher order approximations, yet those can quickly go outside of the permissible range dictated by the production theory axioms, e.g., requiring monotonicity of the production function (or free disposability of the related technology set). Moreover, imposing monotonicity together with convexity on the higher order approximations usually makes them much less different from the first order approximations anyway.

[33] All the computations were done in Matlab using codes programmed by the author, involving standard Matlab library and some codes (for functions for estimation of DEA and FDH) are adopted from Matlab codes of Leopold Simar.

Usually, in MC studies researchers choose particular coefficients for x and y in (36). Here, we would like to avoid limiting the generality of results to any particular coefficients and so we chose them randomly in each MC replication. We used the following procedure: for each $l \in \{2, ..., N\}$ we set $\alpha_l = \left(\widetilde{\alpha}_l / \left(\sum_{l=2}^{N} \widetilde{\alpha}_l\right)\right)(1 - \alpha_1)$ where $\widetilde{\alpha}_l \overset{iid}{\sim} \text{Uniform}(0, 1)$, while α_1 is normalized to some arbitrary value in $(0, 1)$, e.g. 0.25 as per the results below. Meanwhile, for each $m \in \{1, ..., M\}$ we set $\beta_m = \widetilde{\beta}_m / \sum_{m=1}^{M} \widetilde{\beta}_m$ where $\widetilde{\beta}_m \overset{iid}{\sim} \text{Uniform}(0, 1)$. These coefficients are assumed to be *unknown* to a researcher and so will not be used to reduce the dimensions of inputs and outputs for DEA, but only for generating the data.

Generating random data for reasonably general multi-input-multi-output technologies is a challenging task. There are different ways to randomly generate inputs and outputs that will obey (36). The approach we take here is to first generate efficient outputs for each $k \in \{1, ..., n\}$, e.g., as $\widetilde{y}_m^k \overset{iid}{\sim} \text{Uniform}(0.1, 1)$ for each $m \in \{1, ..., M\}$, and then generate $N - 1$ of inputs, e.g., as $x_l^k \overset{iid}{\sim} \text{Uniform}(0.1, 1)$ for each $l \in \{2, ..., N\}$, and then define the remaining input, x_1, in terms of the other generated inputs and all the outputs via (36), i.e., for each $k \in \{1, ..., n\}$, set

$$x_1^k := \left(\left(\left(\sum_{m=1}^{M} \beta_m (\widetilde{y}_m^k)^2\right)^{1/2}\right)^{1/\gamma} / \left(\prod_{l=2}^{N} (x_l^k)^{\alpha_l}\right)\right)^{1/\alpha_1} \tag{37}$$

Now, the reason for setting α_1 to some fixed value should be clear—when uniform gives a value close to zero then $1/\alpha_1$ will be very large and since it is the exponent in the power it will make x_1^k extremely large if the value inside the parentheses is greater than 1 or nearly zero otherwise, causing computational problems. We avoid such problems by this normalization, without losing generality of the argument (because the order of elements in x is arbitrary and all the random values are generated in every replication). This connects all inputs and all outputs in a similar way as a production function would, but in a multi-dimensional case and ensuring convexity and monotonicity properties.

Furthermore, we assume that a researcher observes all original inputs but instead of the maximal or efficient output $\tilde{y}^k$ the researcher observes inefficient outputs defined as

$$y^k = \widetilde{y}^k \times \exp(-u^k), \tag{38}$$

where $u^k \overset{iid}{\sim} |N(0, \sigma_u^2)|$, for $k = 1, ..., n$, and so i.e., $\exp(-u^k)$ is the inefficiency of DMU k. The sample of observations available to a researcher at a replication is $\mathcal{S}_n = \{(x^k, y^k)\}_{k=1}^n$, which is randomly generated at every replication.

Our interest is in the cases when $N \geq n$ or $M \geq n$ (or both) and so, at least in relative terms (dimension relative to the sample size), the described data generating process poses a 'big data' problem for the DEA and FDH, because they cannot be relied upon when using too many inputs or too many outputs, especially when $N \geq n$ or $M \geq n$.

In all the tables below we set $N = M = 100$ and $n \in \{20, 50, 100\}$ and then consider several cases of aggregation of 100 inputs and 100 outputs into a smaller number of inputs and outputs.[34]The idea is to see how close we get to the true efficiency scores, despite such aggregation of many inputs or outputs into just a few. Note that we deliberately look at relatively small samples, under 100, because that is what many applications used and will require to use and we want to get a sense through these simulations of how serious the errors of estimation are likely to be for such studies.

Importantly, note that the use of PCA could be of little help for the described scenarios because (even if there were much less inputs and outputs) the inputs are generated from uniform distribution and so, by construction, there is a very low empirical correlation between the inputs (and zero theoretical correlation). This scenario is therefore

[34]We also tried many other combinations of N, M and n including $N = M \geq 1000$, as could be the case in some 'big data' that contains characteristics of individuals whose performance is to be analyzed; the results were qualitatively the same and even similar quantitatively.

particularly important to understand the value added of the price-based aggregation of inputs and outputs relative to the correlation-based aggregation such as PCA, etc.

4.2 Different sub-scenarios

As discussed in the previous section, to circumvent the 'big data' problem of the type when the dimension of a production model is too large, especially when $N \geq n$ or $M \geq n$ (or both), one may need to do some aggregation of inputs and of outputs before using DEA or FDH. For the sake of brevity we only focus on the case when all outputs are aggregated into one output, which is a very common case in practice, e.g., when researchers use GDP or total revenue as a proxy for the myriads of outputs. While the same can be justified for the aggregation of inputs, i.e., using total costs as a proxy for the myriads of inputs, sometimes researchers are interested in separately modeling several aggregated inputs that have certain economic meaning, e.g., such as total labor costs, total capital costs, total material costs, etc. Therefore, without much loss of generality, we will consider several scenarios of aggregations of all inputs: (i) into 1 input, (ii) into 2 inputs (the first 50 into one aggregate input and the remaining into the other aggregate input), (iii) into 3 inputs (the first 25 into the first aggregate input, the next 50 into the second aggregate input and the remaining 25 into the third aggregate input), (iv) into 4 inputs (the first 25 into the first aggregate input, the next 25 into the second aggregate input, the next 25 into the third aggregate input and the remaining 25 into the fourth aggregate input).[35] We also tried other levels of aggregation and the results were qualitatively similar.

As discussed in the previous section, a natural way to do the aggregation in such a case is to use input prices, because they contain important information on how the market values those inputs and play a key role in the duality theory in economics. Such aggregation is indeed what is often done by practitioners (or they already get such aggregated data): frequently, all the different types of labor are aggregated into labor costs, all the different types of capital are aggregated into capital costs, all the different types of materials are aggregated into material costs, etc.

We also considered different scenarios for the returns to scale and below we present the results for $\gamma = 1$ and $\gamma = 0.5$. In all the cases we try DEA-CRS, DEA-VRS and DEA-FDH. We also explored many scenarios for σ_u^2 and the results were qualitatively the same (and quantitatively similar) and so we only reported for several cases $\sigma_u = 0.5$, which correspond to $E(\exp(-u^k)) \approx 0.7$.

For each of these scenarios we also distinguished between two important cases:

Case 1. Prices vary across firms and the researcher knows those individual prices and uses them to do the aggregation (or obtained data already aggregated in such a way). For this setup, in the simulations we generated the prices using the following procedure: for each $i = 1, ..., N$ generate $\bar{w}_i \overset{iid}{\sim} \text{Uniform}(0, 1)$ and then for each $k \in \{1, ..., n\}$ we perturb this common price with a uniform noise, within $\pm\epsilon \times 100\%$ difference around the common price. The same procedure is used for generating output prices.

Case 2. Prices vary across firms and a researcher does not know those prices. For this setup, we assume the researcher is able to obtain the sample means of prices for each input and output, based on 15 observations, and imposes the 'Law of one price' assumption to perform the aggregation, based on noisy estimates of prices from a fairly small sample.[36]

For both cases, we considered many choices of $\epsilon \in [0, 0.5]$ and the conclusions were similar and so for the sake of brevity we only present $\epsilon = 0.5$ here.

[35] Note that the order of inputs is largely irrelevant due to the random generation framework we adopted here.

[36] The choice of 15 is motivated by the idea to consider a case of a substantially smaller number than a rule-of-thumb of 30 observations often suggested as a sample size needed to get a reasonable accuracy of estimating a mean from a distribution that is not far from normal.

4.3 Evaluation Criteria

All the results are evaluated by comparing the true efficiency scores to their estimates, using the following criteria: the average bias,[37] the mean absolute deviation (MAD), and the three most popular statistical measures of correlation: Pearson, Spearman and Kendall correlation coefficients (recall that the last two measure rank correlations). The summary tables present the averages of these measures over 1000 replications as well as the corresponding MC-calculated standard deviations over these replications.[38]

4.4 Summary of Results

First, we consider Case 1, i.e., when prices were allowed to vary across DMUs, and they were assumed to be known to the researcher who can use them to do the aggregation (or the researcher obtained already aggregated inputs and outputs).[39] Table 1 presents the sub-case for $\gamma = 1$, $\epsilon = 0.5$, i.e, the true technology is CRS and the price varies uniformly within $\pm 50\%$, which is quite a large degree of variation of prices in practical terms.

One can see that there is indeed always some bias, typically fairly small for DEA-CRS, somewhat larger for DEA-VRS in larger dimensions and much larger for FDH, except for the case with one or two aggregated inputs. Somewhat unexpectedly, the bias for DEA-CRS grows with dimension, although it still remains small. Interestingly, the bias changes its sign depending on the number of aggregated inputs and the prices.

More importantly, note that all the correlation coefficients are very high for DEA (especially DEA-CRS), typically above 0.9, except for Kendall's, which is typically about 0.8. Meanwhile, they are not very high for FDH except for the case with one aggregated input (and one aggregated output). This is in coherence with the theoretical results for the DEA outlined above and the intuition behind the impossibility (at least so far) to establish similarly optimistic results for the FDH.

The best results (for this DGP at least) are the cases when all inputs are aggregated into one or two aggregate proxies for inputs (and one aggregate output): i.e., the increase of disaggregation (or less aggregation) leads to greater bias, due to estimation bias of DEA, which is a manifestation of the 'curse of dimensionality' of DEA and FDH.

Table 2 presents results for the same scenario as in Table 1 except that $\gamma = 0.5$, i.e., technology exhibits decreasing returns to scale. In this case, the DEA-CRS would be a 'misspecified model' if one had the luxury of having enough observations to specify and estimate the true model with such immense dimensions. In reality, when one is using aggregate proxy for many inputs and aggregate proxy for many outputs and the aggregation is done not through the production relationship but using prices (perturbed by noise), estimating the true model is an impossible task and, in general, it is possible that the DEA-CRS estimator could give reasonable approximations, as would DEA-VRS or FDH estimators. Interestingly, this is exactly the case we see in the many simulations, where sometimes for three and four aggregate inputs DEA-CRS outperforms DEA-VRS and certainly FDH, for the relatively small samples. This might be somewhat surprising, yet it is coherent with many claims (often without rigorous proof, at least for a general case) or assumptions that CRS is an appropriate simplifying assumption when data is very aggregated, e.g., as for key macroeconomic variables.

Tables 3 and 4 present results for the same scenarios as those for Tables 1 and 2, respectively, except that it is Case 2, i.e., here prices were allowed to vary across firms and a researcher does not know those prices and is able to obtain the sample means of prices for each input and output, based on 15 observations, and then imposes the 'Law

[37] Note that here the bias is both due to estimation and due to aggregation and so shall be understood in a general sense, as an modeling error.

[38] We also tried many pilot-simulations with 200 replications and several pilot-simulations with 10,000 replications and the results were qualitatively the same (though at much higher computational cost) and even quantitatively very similar, as can be also inferred from the reported standard deviation, and so we proceeded using 1000 replications only for all the variations of the scenarios we presented here.

[39] Here we focus on output orientation since it appears to be the most popular in practice. The results with input orientation are similar. In all the computations, the efficiency scores are transformed to be between 0 and 1 (i.e., to have percentage meaning).

of one price' assumption to perform the aggregation, based on these noisy estimates of prices from a fairly small sample. Interestingly, the estimation error is still fairly small: the bias is relatively small and, most importantly, the correlations are still quite high. Again, the Kendall correlation coefficient is showing a bit of a higher loss than Spearman, yet it is still a relatively small loss considering the impossibility of estimating the disaggregated model.

Overall, this suggests that for DEA relatively little information is lost due to price-based aggregation, even in terms of ranking correlations of DMUs according to the (realized or estimated) efficiency scores. This appears to be a good price to pay for the aggregation of so many dimensions into a few, considering the fact that such aggregation was the key to turn the estimation problem from impossible into possible, preserving substantial amount of valuable information.

5 Concluding Remarks

The first goal was to present a concise review of DEA for the more general Business Analytics community, to bring their attention to the key aspects and developments in DEA. To achieve it, we gave a concise overview, with references to various (although not all) key papers and for reviews/surveys where further details can be found.

The second goal of this paper was to discuss the key aspect (and thus the key challenge) of Business Analytics—the 'big data'—for the context of DEA. We focused on the cases of large dimensions of DEA, and thus complemented the current DEA literature, which mostly focused on the computational aspect of the DEA for 'big data' for cases with large numbers of DMUs (e.g., see Ali (1993), Barr and Durchholz (1997), Dula (2008), Zhu et al. (2018), and most recently Khezrimotlagh et al. (2019)).

The third, and most important, goal of this paper was to discuss possible solutions to one of the major challenges of the 'big data' related to the large dimensions in the context of DEA. To achieve this goal, we supported this discussion with existing theoretical results in the literature (although some of them originally were not cast in the 'big-data' context). We also presented results of our new simulation study of what seems to be a very promising approach for reducing an immense dimension of data via the price-based aggregation of inputs and outputs. The results from our simulation study suggest that for DEA, such an aggregation indeed leads to some loss of information (as the outlined aggregation theory suggests), yet it appears that this loss is not as large as one might expect, especially considering the fact that such aggregation helps turning the infeasible DEA problem with immense dimensions into a feasible one, with much of useful information preserved and delivered to a researcher.

While we obtained quite optimistic results, as with any other results they should also be taken with a 'grain of salt', realizing that all the discussed approaches have their limitations. Indeed, an obvious limitation of the price aggregation approach we focused on here is that the price information (as any other information) may not be very reliable or even unavailable. Also, as with any other simulation study, the conclusions from our simulations are for the specific scenarios that we considered and so they may not necessarily generalize to every possible case. Indeed, as is usually the case in a MC studies, it might be possible to find some peculiar examples that yield different conclusions. In its turn, this calls for more extensive simulation studies that would consider many more scenarios and cases with different functional forms, different distributions, noise, etc.

Meanwhile, to be more constructive and practical, the following 'recipe' is proposed for DEA practitioners facing the problem of too many inputs or outputs relative to the number of observations:

Step 1. Obtain as raw and as disaggregated data as possible.[40]

Step 2. Consider several ways of aggregating the data:

- Correlation-based approach

[40]Here and at other stages, it is useful to explore the data with various standard statistical tools, to identify potential outliers, to notice potential patterns, etc.

- Price-aggregation approach

Step 3. Understand and acknowledge advantages and limitations of each of these aggregation approaches (which could be specific to the data or the context). Select a preferred approach and a supplementary approach(es) for aggregation of inputs or outputs to obtain manageable DEA models with a relatively small number of inputs and outputs relative to the number of observations.[41]

Step 4. Obtain the estimation results from the preferred approach and a supplementary approach(es), identify any qualitative differences in conclusions between them, analyze them to understand possible explanations for such differences.

Step 5. Present and discuss the results, including the potential controversies due to using different aggregation tactics, listing the limitations and possible extensions for future explorations.

Finally, while the existing tools for dealing with the 'big data' problems are very promising, the search for improving upon them shall continue and shall be encouraged. For example, the Lasso method is one of approaches frequently used in other contexts of 'big data' to reduce dimensions in regression contexts. This approach was introduced in the seminal work of Tibshirani (1996) and may require more attention from the DEA community (e.g., see Lee and Cai (2018) for some recent developments). Among the natural directions for future research, it is also worth mentioning the area of Network-DEA (e.g., see Cook et al. (2010); Kao (2014) for reviews), which may have different specifics on how to reduce large dimensions, whether via correlation-based or price-based approaches or, perhaps, the Lasso method.[42]

[41] It is also important to remember that (as is typically the case in general), there is no panacea on how to reduce the dimensions. What should be selected as a preferred approach or as a supplementary approach can depend on the data or the context, including the so-called out-of-sample information, such as prior knowledge of the industry, expert judgements, etc. What is important is to clearly explain the steps, the priors and the reasoning behind the selected approaches.

[42] We thank the anonymous referee for this insight.

Table 1: Monte-Carlo Results: Case 1, with $\gamma = 1$, $\epsilon = 0.5$.

Est. type	DEA-CRS			DEA-VRS			FDH		
Sample size (n)	20	50	100	20	50	100	20	50	100
One Aggregate Input and One Aggregate Output									
Av. Bias	0.023	0.054	0.074	-0.031	0.006	0.030	-0.048	-0.009	0.017
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Av. MAD	0.061	0.071	0.082	0.066	0.054	0.055	0.078	0.060	0.056
std_{MC}	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.001	0.000
Av. Pearson Corr.	0.943	0.943	0.944	0.875	0.913	0.935	0.855	0.895	0.922
std_{MC}	0.001	0.001	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Spearman Corr.	0.925	0.934	0.940	0.843	0.903	0.932	0.815	0.879	0.915
std_{MC}	0.001	0.001	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Kendall Corr.	0.797	0.794	0.796	0.734	0.780	0.805	0.702	0.750	0.781
std_{MC}	0.002	0.001	0.001	0.003	0.002	0.001	0.004	0.002	0.001
Two Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.010	0.025	0.047	-0.072	-0.023	0.007	-0.115	-0.064	-0.030
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.061	0.060	0.067	0.096	0.068	0.060	0.131	0.095	0.075
std_{MC}	0.001	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.000
Av. Pearson Corr.	0.937	0.936	0.937	0.772	0.841	0.887	0.700	0.777	0.831
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.005	0.003	0.002
Av. Spearman Corr.	0.912	0.924	0.931	0.709	0.817	0.879	0.620	0.728	0.803
std_{MC}	0.001	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.002
Av. Kendall Corr.	0.778	0.778	0.781	0.612	0.698	0.750	0.526	0.609	0.670
std_{MC}	0.002	0.001	0.001	0.004	0.002	0.002	0.005	0.003	0.002
Three Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.038	0.001	0.026	-0.113	-0.056	-0.019	-0.176	-0.126	-0.087
std_{MC}	0.001	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.069	0.058	0.060	0.128	0.088	0.071	0.185	0.142	0.113
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.930	0.931	0.933	0.675	0.766	0.827	0.545	0.644	0.715
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.007	0.004	0.003
Av. Spearman Corr.	0.901	0.916	0.924	0.585	0.716	0.805	0.449	0.555	0.646
std_{MC}	0.002	0.001	0.001	0.006	0.003	0.002	0.007	0.004	0.003
Av. Kendall Corr.	0.764	0.766	0.770	0.504	0.609	0.682	0.381	0.461	0.532
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.003
Four Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.059	-0.019	0.006	-0.150	-0.089	-0.047	-0.224	-0.180	-0.142
std_{MC}	0.001	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.079	0.061	0.058	0.160	0.112	0.086	0.228	0.189	0.157
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.925	0.928	0.931	0.579	0.686	0.766	NaN	0.515	0.599
std_{MC}	0.001	0.001	0.000	0.006	0.004	0.002	NaN	0.004	0.003
Av. Spearman Corr.	0.896	0.911	0.920	0.469	0.608	0.721	NaN	0.405	0.494
std_{MC}	0.002	0.001	0.001	0.006	0.004	0.003	NaN	0.005	0.003
Av. Kendall Corr.	0.759	0.759	0.764	0.405	0.518	0.608	NaN	0.337	0.407
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	NaN	0.004	0.003

Table 2: Monte-Carlo Results: Case 1, with $\gamma = 0.5$, $\epsilon = 0.5$.

Est. type	DEA-CRS			DEA-VRS			FDH		
Sample size (n)	20	50	100	20	50	100	20	50	100
One Aggregate Input and One Aggregate Output									
Av. Bias	0.026	0.057	0.078	-0.030	0.007	0.031	-0.047	-0.008	0.018
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Av. MAD	0.064	0.074	0.086	0.066	0.054	0.056	0.078	0.060	0.056
std_{MC}	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.001	0.000
Av. Pearson Corr.	0.939	0.939	0.940	0.872	0.913	0.935	0.851	0.895	0.922
std_{MC}	0.001	0.001	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Spearman Corr.	0.921	0.930	0.936	0.841	0.903	0.932	0.812	0.880	0.916
std_{MC}	0.001	0.001	0.000	0.004	0.002	0.001	0.004	0.002	0.001
Av. Kendall Corr.	0.791	0.787	0.789	0.732	0.780	0.805	0.700	0.750	0.782
std_{MC}	0.002	0.001	0.001	0.003	0.002	0.001	0.004	0.002	0.001
Two Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.008	0.028	0.051	-0.070	-0.021	0.009	-0.111	-0.061	-0.026
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.062	0.063	0.070	0.095	0.068	0.060	0.129	0.093	0.075
std_{MC}	0.001	0.000	0.001	0.001	0.000	0.000	0.001	0.001	0.000
Av. Pearson Corr.	0.933	0.932	0.933	0.773	0.844	0.886	0.702	0.779	0.833
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.006	0.003	0.002
Av. Spearman Corr.	0.909	0.920	0.927	0.712	0.820	0.879	0.624	0.732	0.807
std_{MC}	0.001	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.002
Av. Kendall Corr.	0.773	0.771	0.774	0.614	0.700	0.750	0.530	0.612	0.674
std_{MC}	0.002	0.001	0.001	0.004	0.002	0.002	0.005	0.003	0.002
Three Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.035	0.004	0.029	-0.109	-0.053	-0.016	-0.171	-0.120	-0.080
std_{MC}	0.001	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.070	0.061	0.063	0.126	0.087	0.070	0.180	0.138	0.110
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.925	0.926	0.928	0.679	0.768	0.829	0.555	0.652	0.722
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.006	0.004	0.003
Av. Spearman Corr.	0.896	0.911	0.920	0.590	0.720	0.808	0.457	0.565	0.657
std_{MC}	0.002	0.001	0.001	0.006	0.003	0.002	0.007	0.004	0.003
Av. Kendall Corr.	0.756	0.758	0.763	0.508	0.612	0.685	0.388	0.470	0.542
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.003
Four Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.057	-0.017	0.010	-0.147	-0.085	-0.043	-0.218	-0.174	-0.135
std_{MC}	0.001	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.079	0.063	0.060	0.158	0.111	0.085	0.223	0.184	0.152
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.921	0.924	0.926	0.584	0.690	0.769	NaN	0.527	0.606
std_{MC}	0.001	0.001	0.000	0.006	0.003	0.002	NaN	0.004	0.003
Av. Spearman Corr.	0.891	0.906	0.915	0.473	0.614	0.726	NaN	0.417	0.504
std_{MC}	0.002	0.001	0.001	0.006	0.004	0.002	NaN	0.004	0.003
Av. Kendall Corr.	0.752	0.752	0.757	0.409	0.522	0.612	NaN	0.347	0.415
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	NaN	0.004	0.003

Table 3: Monte-Carlo Results: Case 2, with $\gamma = 1$, $\epsilon = 0.5$.

Est. type	DEA-CRS			DEA-VRS			FDH		
Sample size (n)	20	50	100	20	50	100	20	50	100
One Aggregate Input and One Aggregate Output									
Av. Bias	0.011	0.039	0.057	-0.036	-0.002	0.019	-0.053	-0.016	0.007
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Av. MAD	0.050	0.056	0.065	0.062	0.047	0.045	0.074	0.054	0.047
std_{MC}	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000
Av. Pearson Corr.	0.961	0.961	0.962	0.885	0.925	0.949	0.864	0.909	0.937
std_{MC}	0.001	0.000	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Spearman Corr.	0.943	0.953	0.958	0.854	0.914	0.944	0.826	0.892	0.929
std_{MC}	0.001	0.001	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Kendall Corr.	0.831	0.829	0.831	0.755	0.805	0.831	0.722	0.775	0.808
std_{MC}	0.002	0.001	0.001	0.003	0.002	0.001	0.004	0.002	0.001
Two Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.017	0.013	0.034	-0.077	-0.030	-0.002	-0.119	-0.070	-0.037
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.053	0.049	0.053	0.093	0.062	0.051	0.131	0.092	0.070
std_{MC}	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.001
Av. Pearson Corr.	0.956	0.956	0.957	0.782	0.858	0.902	0.701	0.788	0.846
std_{MC}	0.001	0.000	0.000	0.004	0.002	0.001	0.006	0.003	0.002
Av. Spearman Corr.	0.934	0.945	0.951	0.718	0.830	0.892	0.620	0.734	0.817
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.001	0.006	0.004	0.002
Av. Kendall Corr.	0.815	0.814	0.817	0.627	0.721	0.776	0.533	0.622	0.693
std_{MC}	0.002	0.001	0.001	0.004	0.002	0.001	0.005	0.003	0.002
Three Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.041	-0.008	0.015	-0.118	-0.062	-0.027	-0.183	-0.134	-0.094
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.062	0.050	0.048	0.128	0.084	0.064	0.188	0.145	0.112
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.950	0.951	0.953	0.681	0.783	0.845	0.537	0.644	0.725
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.006	0.004	0.003
Av. Spearman Corr.	0.924	0.938	0.946	0.588	0.730	0.820	0.436	0.548	0.652
std_{MC}	0.001	0.001	0.000	0.006	0.003	0.002	0.007	0.004	0.003
Av. Kendall Corr.	0.800	0.802	0.808	0.511	0.629	0.707	0.374	0.460	0.545
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.003
Four Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.059	-0.025	-0.002	-0.154	-0.094	-0.053	-0.228	-0.187	-0.148
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.072	0.054	0.048	0.160	0.110	0.081	0.231	0.193	0.158
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.945	0.949	0.951	0.592	0.701	0.782	NaN	0.512	0.603
std_{MC}	0.001	0.001	0.000	0.006	0.003	0.002	NaN	0.004	0.003
Av. Spearman Corr.	0.920	0.933	0.941	0.479	0.619	0.735	NaN	0.398	0.496
std_{MC}	0.001	0.001	0.000	0.006	0.004	0.002	NaN	0.005	0.003
Av. Kendall Corr.	0.796	0.796	0.801	0.415	0.532	0.629	NaN	0.333	0.412
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	NaN	0.004	0.003

Table 4: Monte-Carlo Results: Case 2, with $\gamma = 0.5$, $\epsilon = 0.5$.

Est. type	DEA-CRS			DEA-VRS			FDH		
Sample size (n)	20	50	100	20	50	100	20	50	100
One Aggregate Input and One Aggregate Output									
Av. Bias	0.014	0.043	0.062	-0.036	-0.000	0.020	-0.052	-0.014	0.009
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001	0.001
Av. MAD	0.053	0.060	0.070	0.062	0.047	0.045	0.075	0.054	0.047
std_{MC}	0.001	0.001	0.001	0.001	0.000	0.000	0.001	0.000	0.000
Av. Pearson Corr.	0.957	0.956	0.958	0.883	0.925	0.948	0.862	0.910	0.936
std_{MC}	0.001	0.000	0.000	0.003	0.002	0.001	0.004	0.002	0.001
Av. Spearman Corr.	0.938	0.948	0.953	0.851	0.914	0.944	0.824	0.893	0.929
std_{MC}	0.001	0.001	0.000	0.004	0.002	0.001	0.004	0.002	0.001
Av. Kendall Corr.	0.821	0.819	0.822	0.752	0.805	0.831	0.719	0.775	0.808
std_{MC}	0.002	0.001	0.001	0.003	0.002	0.001	0.004	0.002	0.001
Two Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.015	0.017	0.039	-0.074	-0.027	0.000	-0.114	-0.065	-0.032
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.055	0.052	0.058	0.092	0.061	0.052	0.126	0.089	0.068
std_{MC}	0.001	0.000	0.000	0.001	0.000	0.000	0.001	0.001	0.001
Av. Pearson Corr.	0.952	0.951	0.952	0.784	0.861	0.903	0.710	0.793	0.849
std_{MC}	0.001	0.000	0.000	0.004	0.002	0.001	0.006	0.003	0.002
Av. Spearman Corr.	0.929	0.940	0.946	0.721	0.835	0.894	0.631	0.743	0.822
std_{MC}	0.001	0.001	0.000	0.005	0.002	0.001	0.006	0.004	0.002
Av. Kendall Corr.	0.806	0.804	0.807	0.631	0.725	0.777	0.542	0.630	0.698
std_{MC}	0.002	0.001	0.001	0.004	0.002	0.001	0.005	0.003	0.002
Three Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.039	-0.004	0.020	-0.113	-0.056	-0.022	-0.174	-0.123	-0.083
std_{MC}	0.001	0.001	0.001	0.001	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.064	0.052	0.052	0.124	0.082	0.063	0.180	0.137	0.105
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.944	0.945	0.947	0.683	0.787	0.848	0.551	0.660	0.738
std_{MC}	0.001	0.001	0.000	0.005	0.003	0.002	0.006	0.004	0.003
Av. Spearman Corr.	0.917	0.931	0.939	0.593	0.739	0.827	0.452	0.568	0.673
std_{MC}	0.001	0.001	0.000	0.006	0.003	0.002	0.007	0.004	0.003
Av. Kendall Corr.	0.789	0.791	0.796	0.516	0.636	0.713	0.388	0.477	0.562
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	0.006	0.004	0.003
Four Aggregate Inputs and One Aggregate Output									
Av. Bias	-0.057	-0.022	0.003	-0.148	-0.087	-0.047	-0.220	-0.177	-0.136
std_{MC}	0.001	0.001	0.001	0.002	0.001	0.001	0.002	0.001	0.001
Av. MAD	0.073	0.056	0.051	0.155	0.106	0.078	0.223	0.184	0.149
std_{MC}	0.001	0.000	0.000	0.001	0.001	0.000	0.002	0.001	0.001
Av. Pearson Corr.	0.940	0.943	0.945	0.597	0.711	0.788	NaN	0.533	0.621
std_{MC}	0.001	0.001	0.000	0.006	0.003	0.002	NaN	0.004	0.003
Av. Spearman Corr.	0.914	0.928	0.935	0.485	0.632	0.745	NaN	0.418	0.518
std_{MC}	0.002	0.001	0.000	0.006	0.004	0.002	NaN	0.005	0.003
Av. Kendall Corr.	0.786	0.785	0.790	0.421	0.543	0.639	NaN	0.350	0.431
std_{MC}	0.002	0.001	0.001	0.005	0.003	0.002	NaN	0.004	0.003

References

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):260 – 273. Data Envelopment Analysis.

Adler, N. and Yazhemsky, E. (2010). Improving discrimination in data envelopment analysis: Pca-dea or variable reduction. *European Journal of Operational Research*, 202(1):273–284. cited By 101.

Afriat, S. N. (1972). Efficiency estimation of production functions. *International Economic Review*, 13(3):568–598.

Ali, A. I. (1993). Streamlined computation for data envelopment analysis. *European Journal of Operational Research*, 64(1):61 – 67.

Allen, R., Athanassopoulos, A., Dyson, R. G., and Thanassoulis, E. (1997). Weights restrictions and value judgements in data envelopment analysis: Evolution, development and future directions. *Annals of Operations Research*, 73:13–34.

Bădin, L., Daraio, C., and Simar, L. (2010). Optimal bandwidth selection for conditional efficiency measures: A data-driven approach. *European Journal of Operational Research*, 201(2):633–640.

Bădin, L., Daraio, C., and Simar, L. (2012). How to measure the impact of environmental factors in a nonparametric production model. *European Journal of Operational Research*, 223(3):818–833.

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

Balk, B. M. (2004). Decompositions of Fisher indexes. *Economics Letters*, 82(1):107–113.

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

Balk, B. M. and Diewert, W. E. (2001). A characterization of the Törnqvist price index. *Economics Letters*, 72(3):279–281.

Banker, R. D. (1993). Maximum likelihood, consistency and data envelopment analysis: A statistical foundation. *Management Science*, 39(10):1265–1273.

Banker, R. D., Charnes, A., and Cooper, W. W. (1984). Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management Science*, 30(9):1078–1092.

Banker, R. D. and Morey, R. C. (1986a). Efficiency analysis for exogenously fixed inputs and outputs. *Operations Research*, 34(4):513–521.

Banker, R. D. and Morey, R. C. (1986b). The use of categorical variables in data envelopment analysis. *Management Science*, 32(12):1613–1627.

Banker, R. D. and Natarajan, R. (2008). Evaluating contextual variables affecting productivity using data envelopment analysis. *Operations Research*, 56(1):48–58.

Barr, R. and Durchholz, M. (1997). Parallel and hierarchical decomposition approaches for solving large-scale data envelopment analysis models. *Annals of Operations Research*, 73:339–372. cited By 32.

Barr, R., Seiford, L., and Siems, T. (1994). Forecasting bank failure: A non-parametric frontier estimation approach. *Reserches Economiques de Lovain*, 60(4):417–429.

Beran, R. and Ducharme, G. R. (1991). *Asymptotic Theory for Bootstrap Methods in Statistics*. Montréal, QC: Centre de Recherches Mathématiques.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics*, 9(6):1196–1217.

Blundell, R. and Stoker, T. M. (2005). Heterogeneity and aggregation. *Journal of Economic Literature*, 43(2):347–391.

Bogetoft, P. (1996). DEA on relaxed convexity assumptions. *Management Science*, 42(3):457–465.

Bogetoft, P., Tama, J. M., and Tind, J. (2000). Convex input and output projections of nonconvex production possibility sets. *Management Science*, 46(6):858–869.

Boussofiane, A., Dyson, R. G., and Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, 52(1):1–15.

Caves, D. W., Christensen, L. R., and Diewert, W. E. (1982a). The economic theory of index numbers and the measurement of input, output, and productivity. *Econometrica*, 50(6):1393–1414.

Caves, D. W., Christensen, L. R., and Diewert, W. E. (1982b). Multilateral comparisons of output, input, and productivity using superlative index numbers. *The Economic Journal*, 92(365):73–86.

Cazals, C., Florens, J.-P., and Simar, L. (2002). Nonparametric frontier estimation: A robust approach. *Journal of Econometrics*, 106(1):1–25.

Charnes, A., Cooper, W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444.

Charnes, A., Cooper, W. W., Lewin, A. Y., and Seiford, L. M., editors (1994). *Data Envelopment Analysis: Theory, Methodology, and Applications*. Amsterdam, NL: Springer.

Charnes, A. and Neralic, L. (1990). Sensitivity analysis of the additive model in data envelopment analysis. *European Journal of Operational Research*, 48(3):332 – 341.

Chen, H., Chiang, R., and Storey, V. (2012). Business intelligence and analytics: From big data to big impact. *MIS Quarterly*, 36(4):1165–1188. cited By 1.

Cook, W. and Seiford, L. (2009). Data envelopment analysis (DEA) - thirty years on. *European Journal of Operational Research*, 192(1):1–17.

Cook, W. D., Liang, L., and Zhu, J. (2010). Measuring performance of two-stage network structures by DEA: A review and future perspective. *Omega*, 38(6):423–430.

Cook, W. D., Ruiz, J. L., Sirvent, I., and Zhu, J. (2017). Within-group common benchmarking using dea. *European Journal of Operational Research*, 256(3):901 – 910.

Cook, W. D., Tone, K., and Zhu, J. (2014). Data envelopment analysis: Prior to choosing a model. *Omega*, 44:1 – 4.

Cooper, W. W., Deng, H., Huang, Z., and Li, S. X. (2004). Chance constrained programming approaches to congestion in stochastic data envelopment analysis. *European Journal of Operational Research*, 155(2):487 – 501. Financial Risk in Open Economies.

Cooper, W. W., Huang, Z., Lelas, V., Li, S. X., and Olesen, O. B. (1998). Chance constrained programming formulations for stochastic characterizations of efficiency and dominance in dea. *Journal of Productivity Analysis*, 9(1):53–79.

Cooper, W. W., Seiford, L. M., and Tone, K. (2007). *Data Envelopment Analysis: A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. New York, NY: Springer.

Cooper, W. W., Seiford, L. M., and Zhu, J., editors (2011). *Handbook on Data Envelopment Analysis*. New York, NY: Springer.

Dantzig, G. (1963). *Linear Programming and Extensions*. Princeton, NJ: Princeton University Press.

Dantzig, G. B. (1949). Programming of interdependent activities: Ii mathematical model. *Econometrica*, 7(3):200–211.

Dantzig, G. B. (1955). Optimal solution of a dynamic Leontief model with substitution. *Econometrica*, 23(3):295–302.

Dantzig, G. B. (1982). Reminiscences about the origins of linear programming. *Operations Research Letters*, 1(2):43–48.

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. (2007a). *Advanced Robust and Nonparametric Methods in Efficiency Analysis: Methodology and Applications*. New York, NY: Springer.

Daraio, C. and Simar, L. (2007b). Conditional nonparametric frontier models for convex and nonconvex technologies: A unifying approach. *Journal of Productivity Analysis*, 28(1):13–32.

Daraio, C. and Simar, L. (2014). Directional distances and their robust versions: Computational and testing issues. *European Journal of Operational Research*, 237(1):358–369.

Daraio, C., Simar, L., and Wilson, P. W. (2017). Central limit theorems for conditional efficiency measures and tests of the "separability" condition in nonparametric, two-stage models of production. *The Econometrics Journal*, page Forthcoming.

Daskovska, A., Simar, L., and Van Bellegem, S. (2010). Forecasting the Malmquist productivity index. *Journal of Productivity Analysis*, 33(2):97–107.

Debreu, G. (1951). The coefficient of resource utilization. *Econometrica*, 19(3):273–292.

Deprins, D., Simar, L., and Tulkens, H. (1984). Measuring labour efficiency in post offices. In Marchand, M., Pestieau, P., and Tulkens, H., editors, *The Performance of Public Enterprises: Concepts and Measurement*, pages 243–267. Amsterdam, NL: Springer.

Diewert, E. and Parkan, C. (1983). Linear programming tests of regulatory conditions for production functions. In Eichhorn, R. Henn, K. N. and Shephard, R. W., editors, *Quantitative studies on production and prices*, pages 131–158. Physica Verlag, Wurzburg, West Germany.

Diewert, W. E. (1971). An application of the Shephard duality theorem: A generalized Leontief production function. *Journal of Political Economy*, 79(3):481–507.

Diewert, W. E. (1974). A note on aggregation and elasticities of substitution. *The Canadian Journal of Economics*, 7(1):12–20.

Diewert, W. E. (1976). Exact and superlative index numbers. *Journal of Econometrics*, 4(2):115–145.

Diewert, W. E. (1978). Superlative index numbers and consistency in aggregation. *Econometrica*, 46(4).

Diewert, W. E. (1980). *Aggregation Problems in the Measurement of Capital*, pages 433–528. Chicago, IL: University of Chicago Press.

Diewert, W. E. (1981). The economic theory of index numbers: A survey. In Deaton, A., editor, *Essays in the Theory and Measurement of Consumer Behaviour (in Honour of Richard Stone)*, pages 163–208. London, UK: Cambridge University Press.

Diewert, W. E. (1983). The measurement of waste within the production sector of an open economy. *The Scandinavian Journal of Economics*, 85(2):159–179.

Diewert, W. E. (1992a). Fisher ideal output, input, and productivity indexes revisited. *Journal of Productivity Analysis*, 3(3):211–248.

Diewert, W. E. (1992b). The measurement of productivity. *Bulletin of Economic Research*, 44(3):163–198.

Diewert, W. E. and Nakamura, A. O., editors (1993). *Essays in Index Number Theory*, volume 1. Amsterdam, NL: Elsevier.

Dula, J. (2008). A computational study of dea with massive data sets. *Computers and Operations Research*, 35(4):1191–1203. cited By 31.

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*, 132:245–259.

Dyson, R. G. and Thanassoulis, E. (1988). Reducing weight flexibility in data envelopment analysis. *Journal of the Operational Research Society*, 39(6):563–576.

Efron, B. (1979). Another look at the jackknife. *Annals of Statistics*, 7(1):1–26.

Emrouznejad, A., Parker, B., and Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. *Journal of Socio-Economics Planning Science*, 42(3):151–157.

Färe, R. and Grosskopf, S. (1985). A nonparametric cost approach to scale efficiency. *The Scandinavian Journal of Economics*, 87(4):594–604.

Färe, R., Grosskopf, S., and Logan, J. (1983). The relative efficiency of Illinois electric utilities. *Resources and Energy*, 5(4):349–367.

Färe, R., Grosskopf, S., Norris, M., and Zhang, Z. (1994). Productivity growth, technical progress, and efficiency change in industrialized countries. *American Economic Review*, 84(1):66–83.

Färe, R., Grosskopf, S., and Zelenyuk, V. (2004). Aggregation bias and its bounds in measuring technical efficiency. *Applied Economics Letters*, 11(10):657–660.

Färe, R. and Primont, D. (1995). *Multi-Output Production and Duality: Theory and Applications*. New York, NY: Kluwer Academic Publishers.

Färe, R. and Zelenyuk, V. (2002). Input aggregation and technical efficiency. *Applied Economics Letters*, 9(10):635–636.

Farrell, M. J. (1957). The measurement of productive efficiency. *Jornal of the Royal Statistical Society. Series A (General)*, 120(3):253–290.

Fisher, I. (1921). The best form of index number. *Journal of the American Statistical Association*, 17(133):533–537.

Fisher, I. (1922). *The Making of Index Numbers*. Boston, MA: Houghton Mifflin.

Førsund, F. R. and Sarafoglou, N. (2002). On the origins of data envelopment analysis. *Journal of Productivity Analysis*, 17(1):23–40.

Frisch, R. (1930). Necessary and sufficient conditions regarding the form of an index number which shall meet certain of Fisher's tests. *American Statistical Association Journal*, 25(172):397–406.

Frisch, R. (1936). Annual survey of general economic theory: The problem of index numbers. *Econometrica*, 4(1):1–38.

Georgescu-Roegen, N. (1951). The aggregate linear production function and its applications to von Neumann's economic model. In Koopmans, T., editor, *Activity Analysis of Production and Allocation*. New York, NY: Wiley.

Gijbels, I., Mammen, E., Park, B. U., and Simar, L. (1999). On estimation of monotone and concave frontier functions. *Journal of the American Statistical Association*, 94(445):220–228.

Gorman, W. M. (1953). Community preference fields. *Econometrica*, 21(1):63–80.

Gorman, W. M. (1959). Separable utility and aggregation. *Econometrica*, 27(3):469–481.

Hall, P. (1992). On bootstrap confidence intervals in nonparametric regression. *Annals of Statistics*, 20(2):695–711.

Härdle, W. and Simar, L. (2015). *Applied Multivariate Statistical Analysis*. Springer Berlin Heidelberg.

Jeong, S.-O. and Simar, L. (2006). Linearly interpolated FDH efficiency score for nonconvex frontiers. *Journal of Multivariate Analysis*, 97(10):2141–2161.

Kantorovich, L. V. (1960). Mathematical methods of organizing and planning production. *Management Science*, 6(4):366–422.

Kao, C. (2014). Network data envelopment analysis: A review. *European Journal of Operational Research*, 239(1):1–16.

Khezrimotlagh, D., Zhu, J., Cook, W. D., and Toloo, M. (2019). Data envelopment analysis and big data. *European Journal of Operational Research*, 274(3):1047 – 1054.

Kneip, A., Park, B. U., and Simar, L. (1998). A note on the convergence of nonparametric DEA estimators for production efficiency scores. *Econometric Theory*, 14(6):783–793.

Kneip, A., Simar, L., and Wilson, P. W. (2008). Asymptotics and consistent bootstraps for DEA estimators in nonparametric frontier models. *Econometric Theory*, 24(6):1663–1697.

Kneip, A., Simar, L., and Wilson, P. W. (2015). When bias kills the variance: Central limit theorems for DEA and FDH efficiency scores. *Econometric Theory*, 31(2):394–422.

Kneip, A., Simar, L., and Wilson, P. W. (2016). Testing hypotheses in nonparametric models of production. *Journal of Business & Economic Statistics*, 34(3):435–456.

Kohavi, R., Rothleder, N., and Simoudis, E. (2002). Emerging trends in business analytics. *Communications of the ACM*, 45(8):45–48. cited By 124.

Konüs, A. A. (1939 [1924]). The problem of the true index of the cost of living. *Econometrica*, 7(1):10–29. Translated into English and published in 1939.

Konüs, A. A. and Byushgens, S. S. (1926). K probleme pokupatelnoi sily deneg. *Voprosy Konyunktury*, 2(1):151–172. English translation of Russian title: "On the problem of the purchasing power of money".

Koopmans, T. (1951a). *Activity Analysis of Production and Allocation*. New York, NY: Wiley.

Koopmans, T. C. (1951b). Analysis of production as an efficienct combination of activities. *Activity Analysis of Production and Allocation*, 13:33–37.

Korostelev, A., Simar, L., and Tsybakov, A. B. (1995a). Efficient estimation of monotone boundaries. *Annals of Statistics*, 23(2):476–489.

Korostelev, A., Simar, L., and Tsybakov, A. B. (1995b). On estimation of monotone and convex boundaries. *Publications de l'Institut de Statistique des Universités de Paris*, 39(1):3–18.

Kumar, S. and Russell, R. R. (2002). Technological change, technological catch-up, and capital deepening: Relative contributions to growth and convergence. *American Economic Review*, 92(3):527–548.

Kumbhakar, S. C., Park, B. U., Simar, L., and Tsionas, E. G. (2007). Nonparametric Stochastic Frontiers: A Local Maximum Likelihood Approach. *Journal of Econometrics*, 137(1):1–27.

Laspeyres, E. (1871). Die berechnmung einer mittleren waarenpreissteigerung. *Jahrbücher für Nationalökonomie und Statistik*, 16:296–314.

Lee, C.-Y. and Cai, J.-Y. (2018). Lasso variable selection in data envelopment analysis with small datasets. *Omega*.

Leontief, W. (1925). Die bilanz der russischen volkswirtschaft - eine methodologische untersuchung. *Weltwirtschaftliches Archiv*, 22(2):338–344.

Liberatore, M. and Luo, W. (2011). Informs and the analytics movement: The view of the membership. *Interfaces*, 41(6):578–589.

Lustig, I., Dietrich, B., Johnson, C., and Dziekan, C. (2010). The analytics journey. *Analytics Magazine*, November/December:11–13.

Marr, B. (2015). *Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance*. Wiley.

Misiunas, N., Oztekin, A., Chen, Y., and Chandra, K. (2016). Deann: A healthcare analytic methodology of data envelopment analysis and artificial neural networks for the prediction of organ recipient functional status. *Omega*, 58:46 – 54.

Mortenson, M. J., Doherty, N. F., and Robinson, S. (2015). Operational research from taylorism to terabytes: A research agenda for the analytics age. *European Journal of Operational Research*, 241(3):583 – 595.

Olesen, O. B. and Petersen, N. C. (1995). Chance constrained efficiency evaluation. *Management Science*, 41(3):442–457.

Olesen, O. B. and Petersen, N. C. (2016). Stochastic data envelopment analysis–a review. *European Journal of Operational Research*, 251(1):2 – 21.

Paasche, H. (1874). Über die preisentwicklung der letzten jahre nach den hamburger börsennotirungen. *Jahrbücher für Nationalökonomie und Statistik*, 23:168–178.

Park, B. U., Jeong, S.-O., and Simar, L. (2010). Asymptotic distribution of conical-hull estimators of directional edges. *Annals of Statistics*, 38(3):1320–1340.

Park, B. U., Simar, L., and Weiner, C. (2000). The FDH estimator for productivity efficiency scores: Asymptotic properties. *Econometric Theory*, 16(6):855–877.

Park, B. U., Simar, L., and Zelenyuk, V. (2008). Local likelihood estimation of truncated regression and its partial derivatives: Theory and application. *Journal of Econometrics*, 146(1):185–198.

Park, B. U., Simar, L., and Zelenyuk, V. (2015). Categorical data in local maximum likelihood: Theory and applications to productivity analysis. *Journal of Productivity Analysis*, 43(2):199–214.

Petersen, N. C. (1990). Data envelopment analysis on a relaxed set of assumptions. *Management Science*, 36(3):305–314.

Ranyard, J., Fildes, R., and Hu, T.-I. (2015). Reassessing the scope of or practice: The influences of problem structuring methods and the analytics movement. *European Journal of Operational Research*, 245(1):1 – 13.

Ray, S. C. (2004). *Data Envelopment Analysis: Theory and Techniques for Economics and Operations Research*. New York, NY: Cambridge University Press.

Ruggiero, J. (2005). Impact assessment of input omission on dba. *International Journal of Information Technology and Decision Making*, 4(3):359–368. cited By 40.

Seitz, W. D. (1970). The measurement of efficiency relative to a frontier production function. *American Journal of Agricultural Economics*, 52(4):505–511.

Seitz, W. D. (1971). Productive efficiency in the steam-electric generating industry. *Journal of Political Economy*, 79(4):878–86.

Sengupta, J. K. (1998). Stochastic data envelopment analysis: a new approach. *Applied Economics Letters*, 5(5):287–290.

Shephard, R. W. (1953). *Cost and Production Functions*. Princeton, NJ: Princeton University Press.

Sherman, D. and Zhu, J. (2013). Analyzing performance of service organizations: Balanced benchmarking can identify best practices that are often hidden. *MIT Sloan Management Review*, 54(4):18–33.

Simar, L. (1992). Estimating efficiencies from frontier models with panel data: A comparison of parametric, non-parametric and semi-parametric methods with bootstrapping. *Journal of Productivity Analysis*, 3(1):171–203.

Simar, L., Van Keilegom, I., and Zelenyuk, V. (2017). Nonparametric least squares methods for stochastic frontier models. *Journal of Productivity Analysis*, 47(3):189–204.

Simar, L., Vanhems, A., and Wilson, P. W. (2012). Statistical inference for DEA estimators of directional distances. *European Journal of Operational Research*, 220(3):853–864.

Simar, L. and Wilson, P. W. (1998). Sensitivity analysis to efficiency scores: How to bootstrap in nonparametric frontier models. *Management Science*, 44(1):49–61.

Simar, L. and Wilson, P. W. (2000a). A general methodology for bootstrapping in non-parametric frontier models. *Journal of Applied Statistics*, 27(6):779–802.

Simar, L. and Wilson, P. W. (2000b). Statistical inference in nonparametric frontier models: The state of the art. *Journal of Productivity Analysis*, 13(1):49–78.

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.

Simar, L. and Wilson, P. W. (2011). Two-stage DEA: Caveat emptor. *Journal of Productivity Analysis*, 36(2):205–218.

Simar, L. and Wilson, P. W. (2013). Estimation and inference in nonparametric frontier models: Recent developments and perspectives. *Foundations and Trends in Econometrics*, 5(3–4):183–337.

Simar, L. and Wilson, P. W. (2015). Statistical approaches for nonparametric frontier models: A guided tour. *International Statistical Review*, 83(1):77–110.

Simar, L. and Zelenyuk, V. (2006). On testing equality of distributions of technical efficiency scores. *Econometric Reviews*, 25(4):497–522.

Simar, L. and Zelenyuk, V. (2007). Statistical inference for aggregates of Farrell-type efficiencies. *Journal of Applied Econometrics*, 22(7):1367–1394.

Simar, L. and Zelenyuk, V. (2011). Stochastic FDH/DEA estimators for frontier analysis. *Journal of Productivity Analysis*, 36(1):1–20.

Simar, L. and Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research*, 166(1):139–149.

Stoker, T. M. (1993). Empirical approaches to the problem of aggregation over individuals. *Journal of Economic Literature*, 31(4):1827–1874.

Tauer, L. W. (2001). Input aggregation and computed technical efficiency. *Applied Economics Letters*, 8:295–297.

Theil, H. (1954). *Linear aggregation of economic relations.* Contributions to economic analysis. North-Holland Pub. Co.

Thomas, A. C. and Tauer, L. W. (1994). Linear input aggregation bias in nonparametric technical efficiency measurement. *Canadian Journal of Agricultural Economics/Revue canadienne d'agroeconomie*, 42(1):77–86.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Törnqvist, L. (1936). The Bank of Finland's consumption price index. *Bank of Finland Monthly Bulletin*, 16(10):1–8.

Ueda, T. and Hoshiai, Y. (1997). Application of principal component analysis for parsimonious summarization of dea inputs and/or outputs. *Journal of the Operations Research Society of Japan*, 40(4):466–478.

von Neumann, J. (1945). A model of general equilibrium. *Review of Economic Studies*, 13(1):1–9.

Wheelock, D. and Wilson, P. (1995). Explaining bank failures: Deposit insurance, regulation, and efficiency. *The Review of Economics and Statistics*, 77(4):689–700.

Wheelock, D. C. and Wilson, P. W. (2000). Why do Banks Disappear? The Determinants of U.S. Bank Failures and Acquisitions. *The Review of Economics and Statistics*, 82(1):127–138.

Wilson, P. W. (2018). Dimension reduction in nonparametric models of production. *European Journal of Operational Research*, 267(1):349 – 367.

Zhu, Q., Wu, J., and Song, M. (2018). Efficiency evaluation based on data envelopment analysis in the big data context. *Computers Operations Research*, 98:291 – 300.