Centre for Efficiency and Productivity Analysis

Working Paper Series
No. WP11/2020

Aggregation of Outputs and Inputs for DEA Analysis of Hospital Efficiency: Economics, Operations Research and Data Science Perspectives

Bao Hoang Nguyen and Valentin Zelenyuk

Date: December 2020

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

ISSN No. 1932 - 4398

Aggregation of Outputs and Inputs for DEA Analysis of Hospital Efficiency: Economics, Operations Research and Data Science Perspectives

Bao Hoang Nguyen[*] Valentin Zelenyuk[†]

December 18, 2020

Abstract

Data envelopment analysis (DEA) has been widely recognised as a powerful tool for performance analysis over the last four decades. The application of DEA in empirical works, however, has become more challenging, especially in the modern era of big data, due to the so-called 'curse of dimensionality'. Dimension reduction has been recently considered as a useful technique to deal with the 'curse of dimensionality' in the context of DEA with large dimensions for inputs and outputs. In this study, we investigate the two most popular dimension reduction approaches: PCA-based aggregation and price-based aggregation for hospital efficiency analysis. Using data on public hospitals in Queensland, Australia, we find that the choice of price systems (with small variation in prices) does not significantly affect the DEA estimates under the price-based aggregation approach. Moreover, the estimated efficiency scores from DEA models are also robust with respect to the two different aggregation approaches.

Keywords: Hospital efficiency, big wide data, DEA, PCA-based aggregation, price-based aggregation

JEL Codes: C24, C61, I11, I18.

[*] School of Economics, University of Queensland, Brisbane, Qld 4072, Australia

[†] School of Economics and Centre for Efficiency and Productivity Analysis, University of Queensland, Brisbane, Qld 4072, Australia

1 Introduction

Envelopment estimators in general and data envelopment analysis (DEA) in particular, have been widely recognised as a powerful tool for performance analysis.[1] Over the four decades since the seminal work of Charnes et al. (1978), DEA has been applied to study the efficiency and productivity of decision making units (DMUs) in various economic sectors. Among these, the healthcare industry, especially hospitals, has been one of the most active sectors of research, where many state-of-the-art developments of DEA have been utilised to provide empirical insights into a large number of papers (e.g., see reviews in Hollingsworth, 2003, 2008; O'Neill et al., 2008; Kohl et al., 2019).

Despite its popularity, the application of DEA in empirical works has been more challenging, especially in the modern era of analytics, where big data becomes the norm. The challenge comes from the well-know fact that DEA, as with virtually any nonparametric estimator, suffers from the 'curse of dimensionality'. That is the accuracy and the discrimination power of DEA decrease when the dimension of input-output space increases. The issue is even worse in the context of big wide data, where the number of inputs and outputs is more than the number of observations: DEA becomes practically infeasible!

Interestingly, the challenge of high dimensionality has been there in the analysis of hospital efficiency long before the wave of modern big data. Since its first application in the U.S. in the early 1980s, hospital product in many countries has been defined using the diagnosis related group (DRG) classification scheme (see more detailed discussions in Fetter, 1991). Under the classification system, hospital product is classified into hundreds of DRGs, grouping together patients with similar diagnoses who require similar hospitals services. Each of these hundreds of DRGs represents a separate type of output of hospitals and counts into the dimension of the production space, which usually turns out to be larger than the typical sample sizes in many applied works in the field.

To make DEA models feasible, a natural and common approach in the literature is to use the so-called casemix weighted episodes (or discharges or patient days). A casemix weighted episode is a weighted sum of the number of episodes in each DRG with the weight being the DRG's casemix cost weight, which can be viewed as the relative price of each DRG in the funding systems. This approach is actually an example of the price-based aggregation that is one of the most powerful techniques for dimension reduction to

[1] Stochastic frontier analysis (SFA) (Aigner et al., 1977; Meeusen & van Den Broeck, 1977) is another popular method for productivity and efficiency analysis. Recently, more advanced methodologies have been developed in the literature such as stochastic DEA (Simar & Zelenyuk, 2011), order-m frontiers (Cazals et al., 2002), order-α quantile frontiers (Aragon et al., 2005; Daouia & Simar, 2007), to mention a few.

deal with the 'curse of dimensionality' in the context of DEA with big wide data.[2] The other powerful technique is the principle component analysis (PCA)-based aggregation[3], which is based on the eigendecomposition of the moment (or correlation) matrix of inputs or outputs or their subsets.[4] The performance of these two techniques in the context of DEA with big data have recently been investigated in the literature, with intensive Monte Carlo simulations therein showing that the gain of dimension reduction is usually more than the information lost (e.g., see Wilson, 2018; Zelenyuk, 2020).

Back to the case of output aggregation in hospital efficiency analysis, some questions are of interest, but are usually overlooked by DEA practitioners, and now may be a good time to revisit these given the increasing attention to DEA as "data enabled analytics" under big data (Zhu, 2020). In particular, some of these questions include: (i) For price-based aggregation, what weights should be used, e.g., a constant weight system or a weight system varying across years? (This question is typical relevant to the case of short panel data when researchers usually pool data across years to estimate a grand frontier)? (ii) What about PCA-based aggregation, i.e., are the estimated efficiency scores robust with respect to the different dimension reduction methods?

In this study, we examine these empirical questions using a fairly aggregated dataset of public hospitals in Queensland, Australia, which, among other things, consists of 153,472 data points for 704 different types of inpatient procedures across 109 hospitals in two years. We find that with a small variation in prices across years, the choice of price (weight) systems does not significantly effect the estimated efficiency scores. Moreover, the estimated efficiency scores from the DEA models are also robust with respect to the two different aggregation approaches. The robustness of the results suggests that PCA-based aggregation can be viewed as a viable alternative for DEA practitioners who are unable to/or unwilling to use the price-based approach, e.g., due to unavailable or unreliable price information.

The rest of the paper is structured as follows. Section 2 discusses methodologies including envelopment estimators and aggregation methods. Section 3 briefly reviews the statistical methods to analyse the efficiency scores. Section 4 describes data and variables used in the study. Section 5 compares the estimated efficiency scores among different aggregation approaches. Section 6 provides concluding remarks.

[2] E.g., see Zelenyuk (2020) for evidence about the performance of price-based aggregation for dimension reduction in the context of DEA with big data.

[3] E.g., see Wilson (2018) for evidence about the performance of PCA-based aggregation for dimension reduction in the context of DEA with big data.

[4] Another promising technique that has been recently applied into the DEA context for dimension reduction is Least Absolute Shrinkage and Selection Operator (LASSO) (see more discussions in Chen et al., 2020; Lee & Cai, 2020, and references therein).

2 Methodology

2.1 Envelopment estimators

Let $\mathcal{X}^n = \left\{ \left(x^k, y^k\right) \mid k = 1, \cdots, n \right\}$ be a set of input-output allocations of a sample of n DMUs, where $x^k = \left(x_1^k, \cdots, x_N^k\right) \in \Re_+^N$ is an $N-$dimensional vector of inputs and $y^k = \left(y_1^k, \cdots, y_M^k\right) \in \Re_+^M$ is an $M-$dimensional vector of outputs.[5]

Under the assumption of Constant Returns to Scale (CRS), the DEA-estimator of the Debreu-Farell output oriented technical efficiency score of a DMU with input-output allocation of (x^j, y^j) is given by [6]

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

One can add an additional constraint, $\sum_{k=1}^{n} z^k = 1$, into equation (1) to obtain the DEA-estimator under the assumption of Variable Returns to Scales (VRS), specifically

$$\begin{aligned}
\widehat{OTE}_{DEA-VRS}\left(x^j, y^j | \mathcal{X}^n\right) \equiv \max_{\theta, z^1, \cdots, z^n} & \ \theta \\
\text{s.t.} & \\
& \sum_{k=1}^{n} z^k y_m^k \geq \theta y_m^j, \ m = 1, \cdots, M, \\
& \sum_{k=1}^{n} z^k x_l^k \leq x_l^j, \ l = 1, \cdots, N, \\
& \sum_{k=1}^{n} z^k = 1, \\
& \theta \geq 0, \ z^k \geq 0, \ k = 1, \cdots, n.
\end{aligned} \tag{2}$$

[5] It is worth mentioning here that in this study we pooled data over years to estimate a grand frontier, so n here is total number of observations across years.

[6] There are also other measures of technical efficiency such as the Debreu-Farell input-oriented measure, the Russell measure, the directional distance function, the slack-based measure, etc. Due to the limit of space we focus on the Debreu-Farell output oriented technical efficiency here, and the interested readers can find detailed discussions about the other measures of technical efficiency in Sickles and Zelenyuk (2019, Ch.3 & Ch. 8).

Another popular envelopment estimator is the so-called Free Disposal Hull (FDH) estimator recently developed by Deprins et al. (1984). The FDH estimator of the Debreu-Farell output oriented technical efficiency score can be obtained by solving the following integer program problem

$$\begin{aligned}\widehat{OTE}_{FDH}\left(x^j, y^j|\mathcal{X}^n\right) \equiv \max_{\theta, z^1, \cdots, z^n} & \ \theta \\ & \text{s.t.} \\ & \sum_{k=1}^{n} z^k y_m^k \geq \theta y_m^j,\ m = 1, \cdots, M, \\ & \sum_{k=1}^{n} z^k x_l^k \leq x_l^j,\ l = 1, \cdots, N, \\ & \sum_{k=1}^{n} z^k = 1, \\ & \theta \geq 0,\ z^k \in \{0, 1\},\ k = 1, \cdots, n.\end{aligned} \tag{3}$$

The formulas discussed above are the envelopment estimators of true Debreu-Farell output oriented technical efficiency scores, defined by

$$OTE\left(x, y|\Psi\right) \equiv \max_{\theta}\left\{\theta \geq 0|\left(x, \theta y\right) \in \Psi\right\}, \tag{4}$$

where Ψ is the set characterising the technology, defined as

$$\Psi \equiv \left\{(x, y) \in \Re_+^N \times \Re_+^M |y \text{ can be produced by } x\right\}. \tag{5}$$

The technology set is assumed to satisfy the standard regularity conditions of the production theory (Shephard, 1953, 1970; Färe & Primont, 1995).

Due to the limit of space we focus here on these most popular DEA models, while there are many other variations of DEA, which also allow modelling the 'undesirable outputs' and 'congesting inputs', incorporating network structure, or imposing various weight restrictions, etc.[7]

2.2 'Curse of dimensionality'

In this section, we will discuss the methodological issues relating to a large dimension of input-output space in the context of envelopment analysis.

As we can see from equations (1) to (3), when the dimension of input-output space $(N+M)$ is larger than sample size (n), the models become practically infeasible. Another issue can also arise when $N + M$ is substantially lower than n, yet still large enough to

[7] See Sickles and Zelenyuk (2019, Ch. 8) for more details and related discussion.

significantly reduce the discrimination power of the models. In such a scenario, the large dimension possibly makes the estimated results not practically useful (i.e., all DMUs may attain an efficiency score of or nearly 100%).[8] More importantly, the accuracy of envelopment estimation depends on the dimension of the input-output space in such a way that the rate of convergence to the truth decreases exponentially when the dimension increases. It is well-known in the literature (e.g., see Kneip et al., 1998) that under appropriate assumptions, the DEA/FDH estimators are consistent with the convergence rate depending on the dimension of input-output space as follows

$$\widehat{OTE}\left(x,y\right) - OTE\left(x,y\right) = O_p\left(n^{-\kappa}\right), \tag{6}$$

where $O_p\left(n^{-\kappa}\right)$ denotes the order of magnitude and κ depends on the estimator. Specifically, we have $\kappa = 2/\left(N+M\right)$ for DEA-CRS estimator, $\kappa = 2/\left(N+M+1\right)$ for DEA-VRS estimator and $\kappa = 1/\left(N+M\right)$ for FDH estimator.

2.3 Aggregation of Inputs and Outputs before DEA

Aggregation of all or some inputs and/or outputs, *inter alia*, can be viewed as a useful method (in some cases, the only method) to deal with the 'curse of dimensionality' in the context of envelopment analysis.[9] For example, in hospital efficiency analysis, labour input is composed of many professionals (e.g., medical officers, nurses, diagnostic and health professionals, etc.). Although it is desirable (and might be feasible in many cases) to include all these labour categories as separate inputs in DEA models to account for the difference in labour composition across hospitals, the estimated efficiency scores from such models might not be reliable due to the issue of the 'curse of dimensionality' as discussed in Section 2.2. On the output side, we also face a similar issue, but more severe. Under the DRG classification system, hospital product is classified into hundreds of DRGs. Each of these hundreds of DRGs represents a separate type of output of hospitals and counts into the dimension of the production space, which usually turns out to be larger than the typical sample sizes in many applied works in the field, making DEA models practically infeasible (e.g., in the working sample of the current study, there are 704 DRGs but only 218 observations).

[8] Also see a related discussion in Charles et al. (2019) who also proposed a novel approach to increase the discriminatory power of DEA in the presence of the 'curse of dimensionality'.

[9] One can also consider the order-α quantile frontier and order-m frontier estimators as alternative approaches for the frontier estimation since these estimators do not suffer from the 'curse of dimensionality'. E.g., see an application of the order-α quantile frontier estimators to hospital data in Nguyen and Zelenyuk (2021b).

In this section we will briefly discuss the two popular approaches to aggregate inputs/outputs which are: (i) PCA-based aggregation and (ii) Price-based aggregation.[10] We will illustrate for the context of the aggregation of some $\hat{M} \leq M$ outputs, and without loss of generality, assume that these outputs are the 1^{st}-output to the $\hat{M}^{th}$-output.[11]

2.3.1 PCA-based aggregation

The idea of PCA-based aggregation in the DEA context is to find linear combinations of inputs/outputs that best represent their variation based on the eigendecomposition of their moment or correlation matrix. The application of PCA in the context of efficiency analysis goes back at least to Zhu (1998), Adler and Golany (2001, 2007), and Mouchart and Simar (2002) and Daraio and Simar (2007). In particular, Zhu (1998) appears to be the first who utilised PCA as a method for evaluating relative performance of DMUs and compared it to DEA. Meanwhile, Adler and Golany (2001, 2007) employed PCA as a dimension reduction method based on the eigendecomposition of the correlation matrix of inputs or outputs. Mouchart and Simar (2002) and Daraio and Simar (2007) also utilised PCA for dimension reduction in the context of DEA, but their approach based on the eigendecomposition of the moment matrix of inputs and outputs. It is worth noting here that PCA-based aggregation using a correlation matrix has an issue in that the first eigenvector (corresponding to the largest eigenvalue) of a correlation matrix may contain both positive and negative values, thus it is not practically useful for the context of envelopment analysis (see more discussions in Wilson, 2018). To avoid this issue, as in Wilson (2018), we follow Mouchart and Simar (2002) and Daraio and Simar (2007) to utilise the PCA-based aggregation using the eigendecomposition of the moment matrix.

Let $\hat{\mathbf{Y}}$ be an $n \times \hat{M}$ matrix stacking the 1^{st}-output to the $\hat{M}^{th}$-output across all n DMUs.[12] $\hat{\mathbf{Y}}$ can be viewed as a cloud of n data points in an $\hat{M}-$dimensional space. The aggregation question here is to find the best subspace of dimension $\tilde{M}$ ($\tilde{M} < \hat{M}$) through the origin to project the cloud onto. Now let us define $\mathcal{B} = \{u_1, \ldots, u_{\tilde{M}}\}$ as an orthonormal basis for the subspace, i.e., $u_{\tilde{m}}$ ($\tilde{m} = 1, \ldots, \tilde{M}$) are all unit vectors and are orthogonal to each other. A natural approach to determine the best subspace is to find the optimal orthonormal basis to solve the following "least square" problem

$$\min_{u_1,\ldots,u_{\tilde{M}}} \sum_{k=1}^{n} \left(\left\| \hat{y}^k \right\|^2 - \left\| p_{\hat{y}^k} \right\|^2 \right),$$

[10] More detailed discussions can be found in Wilson (2018) and Zelenyuk (2020).

[11] Similar description can be given for the case of aggregating many inputs.

[12] Since the efficiency measure is invariant to the units of measurement, it is advised to standardise each column of $\hat{\mathbf{Y}}$ by its standard deviation before the aggregation.

where $\hat{y}^k$ is the k^{th} row of $\hat{\mathbf{Y}}$ and $p_{\hat{y}^k}$ is the projection of $\hat{y}^k$ onto the subspace.

It is well known in mathematics (statistics, optimisation theory, etc.) that the optimal solution for the above problem is the set of orthogonal eigenvectors, $\Lambda_1, \Lambda_2, \ldots \Lambda_{\tilde{M}}$, of $\hat{\mathbf{Y}}^\top \hat{\mathbf{Y}}$, which corresponds to the first $\tilde{M}$ largest eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{\tilde{M}}$ (e.g., see Härdle & Simar, 2020). As in Härdle and Simar (2020), $\Lambda_{\tilde{m}}, \tilde{m} = 1, \ldots, \tilde{M}$, is called the $\tilde{m}-$th factorial axis, and the coordinate of each DMU on the $\tilde{m}-$th factorial axis

$$z^k_{\tilde{m}} = \left(\hat{y}^k\right)^\top \Lambda_{\tilde{m}},\ k = 1, \cdots, n, \tag{7}$$

is called the $\tilde{m}-$th factorial variable. The point $z^k = \left(z^k_1, \ldots, z^k_{\tilde{M}}\right)$ is then the representation of $\hat{y}^k$ in the subspace of dimension $\tilde{M}$. Moreover, the quality of the factorial representations of the original data in the subspace of dimension $\tilde{M}$ can be measured by

$$\delta_{\tilde{M}} := \frac{\sum_{m=1}^{\tilde{M}} \lambda_m}{\sum_{m=1}^{\hat{M}} \lambda_m}. \tag{8}$$

The parameter $\delta_{\tilde{M}}$ represents the percentage of the inertia in the original data that can be explained by the $\tilde{M}$ factorial variables. And as discussed in Daraio and Simar (2007) and Wilson (2018), if the correlation among the $\hat{M}$ outputs is high, then the majority of variation in the data can be explained by the first few factorial variables (i.e., high value of $\delta_{\tilde{M}}$ with a small value of $\tilde{M}$). In this study, we will consider a special case of PCA-based approach, where all $\hat{M}$ outputs are aggregated into an aggregated measure of dimension 1 – the first factorial variable.

2.3.2 Price-based aggregation

Let us assume that all n DMUs face the same output prices for the $\hat{M}$ outputs, denoted as $(p_1, \cdots, p_{\hat{M}}) \in \Re^{\hat{M}}_{++}$. We then can use the output prices as weights to linearly aggregate the $\hat{M}$ outputs. Specifically, in equations (1) to (3), for each of the DMUs, we replace the first $\hat{M}$ outputs by their sub-revenue

$$r^k = \sum_{m=1}^{\hat{M}} p_m y^k_m,\ k = 1, \cdots, n. \tag{9}$$

Besides being very simple, the price-based aggregation approach takes into account the economic valuation of the outputs or inputs (reflected through the prices), which, in principle, might be very different from the weights implied through statistical approaches, such as the PCA-based approach. On the other hand, a limitation of the price-based aggregation is that in practice, some information about the prices might be unavailable.

Under certain conditions, the estimated efficiency scores obtained using the aggregated output are upward biased (compared to the case of using disaggregated outputs), but the bias is bounded by the allocative efficiency. Moreover, it is worth noting here that the assumption of the same output price is for establishing the exact theoretical properties.[13] In practice, when the prices vary to some extent, this price-aggregation approach may still perform well, as confirmed by the Monte Carlo simulations in Zelenyuk (2020). In this study, we also examine the robustness of results with different price systems: same prices vs. varying prices.

3 Statistical Methods to Analyse the Efficiency Scores

After obtaining DEA/FDH estimates of efficiency scores of individual DMUs, researchers usually perform the statistical inference to explore more about the efficiency level of the population or its some groups of interest, for example, comparing the efficiency of different groups in the population by analysing their means, or their densities, or studying the determinants of the efficiency by using regression analysis. Some standard exploratory tools from statistics/econometrics such as the nonparametric statistical tests (e.g., Wilcoxon rank-sum test, Kolmogorov-Smirnov test, etc.), the ordinary least squared regression, or the censored (tobit) regression, were employed to further analyse the efficiency scores from DEA/FDH models.[14] The statistical inference, however, needs to be carefully adapted to the DEA/FDH context given the well-known fact that the DEA/FDH estimates are biased and are serially correlated in a complicated unknown way (Simar & Wilson, 2007, 2015; Sickles & Zelenyuk, 2019).

The adapted methods for statistical inference based on the estimated DEA/FDH efficiency scores have been recently developed and become more popular in the literature thanks to the important works of Simar and Wilson (1998, 2000) and Kneip et al. (2008) (for the bootstrap-based statistical inference for individual efficiency scores), Simar and Zelenyuk (2007) (for the bootstrap–based statistical inference for the weighted mean of efficiency scores), Simar and Zelenyuk (2006) (for the kernel density estimator–based tests), Simar and Wilson (2007) (for the bootstrap truncated regression), Kneip et al. (2015, 2016) and Simar and Wilson (2020) (for the central limit theorems for the simple mean of efficiency scores and related tests), Simar and Zelenyuk (2018) (for the central limit theorems for the weighted mean of efficiency scores), and Simar and Zelenyuk (2020) (for the improvement in finite sample approximation by the central limit theorems), among others.

In this section, we will focus our discussion on the kernel density estimation and related

[13] E.g., see Färe and Grosskopf (1985), Färe et al. (2004) and Zelenyuk (2020) for detailed discussions.

[14] See more discussion about the application as well as the caveats of these methods in the context of DEA/FDH in Grosskopf (1996) and Simar and Wilson (2007).

statistical tests as well as the analysis of simple means and weighted means of estimated efficiency scores, which we will apply in the next section.

3.1 Kernel Density Estimation and Related Tests

To discuss the kernel density estimation, let us denote $\{u_i\}_{i=1}^{n}$ as a set of the realisations of a random variable U and denote f_U as the density function of the variable. Following Rosenblatt (1956), the estimate of f_U at a point u can be obtained by

$$\hat{f}_h(u) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{u_i - u}{h}\right), \tag{10}$$

where $K(\cdot)$ and h are suitable kernel and bandwidth, respectively.

The application of the kernel density estimation for estimating the densities of efficiency scores in the context of DEA/FDH, however, faces two issues. The first issue is the discontinuity problem, that is the probability of having an observation which is fully efficient (i.e., its estimated efficiency score equals one) is always different from zero by construction. The second issue is the problem of bounded support that is the efficiency score is bounded with most of the mass of its distribution being around the boundary, leading to the inconsistency of the kernel density estimator at the boundary.[15] To overcome these issues, one can remove from the sample those observations whose estimated efficiency scores equal to one and then apply the Silverman's (1986) reflection method to estimate the density. Specifically, the kernel density estimate of DEA/FDH efficiency scores at a point u can be obtained by

$$\hat{f}_h^R(u) = \begin{cases} \frac{1}{\tilde{n}h_R}\sum_{i=1}^{\tilde{n}}\left[K\left(\frac{\hat{u}_i - u}{h_R}\right) + K\left(\frac{(2-\hat{u}_i)-u}{h_R}\right)\right],\ u \geq 1 \\ 0,\ \text{otherwise} \end{cases}, \tag{11}$$

where $\{\hat{u}_i\}_{i=1}^{\tilde{n}}$ is a set of estimated output-oriented efficiency scores whose values are greater than unity ($\tilde{n} < n$), h_R is the bandwidth selected for the reflected sample (i.e., $\{\hat{u}_1, ..., \hat{u}_{\tilde{n}}, 2-\hat{u}_i, ..., 2-\hat{u}_{\tilde{n}}\}$).[16] As discuss in Sickles and Zelenyuk (2019), the choice of the kernel $K(\cdot)$ is not very important and one can obtain a good fit when using popular kernels such as Gaussian or Epanechnikov. On the other hand, the selection of the bandwidth is much more important, thus it is advised to use more advanced procedures (e.g., the method of Sheather and Jones (1991), the cross-validation approach, etc.) to select the bandwidth although the Silverman (1986) robust rule-of-thumb bandwidth can give a fairly good fit.

[15] See more discussions in Sickles and Zelenyuk (2019).

[16] The formula can be easily adapted to the case of input-oriented efficiency scores.

In addition to estimating and visualising the density of the estimated efficiency scores, DEA/FDH practitioners are also interested in comparing the densities of efficiency scores of different groups of DMUs in the population using the statistical tests. Here, we will focus our discussion on the Li (1996) test - the test based on the kernel density estimator. To simplify the notations, let us consider only two groups of DMUs, say, group A and Z, with their efficiency realisations and efficiency density function being $\{u_i^A\}_{i=1}^{n_A}$ and $f_A(\cdot)$, $\{u_i^Z\}_{i=1}^{n_Z}$ and $f_Z(\cdot)$, respectively. Formally, the null and alternative hypotheses of the test can be stated as follows

$$\begin{aligned} H_0 :& f_A(u) = f_Z(u), \forall u \text{ in the support of } U^A \text{ and } U^Z \\ H_1 :& f_A(u) \neq f_Z(u), \text{on a set of positive measures.} \end{aligned}$$

The Li (1996) test statistics can be obtained by

$$\hat{L}_{n_A,n_Z,h} = n_A h^{\frac{1}{2}} \hat{J}_{n_A,n_Z,h}/\hat{\sigma}_{\gamma,h} \xrightarrow[\text{under } H_0]{d} N(0,1), \tag{12}$$

where

$$\begin{aligned} \hat{J}_{n_A,n_Z,h} =& \frac{1}{n_A(n_A-1)h} \sum_{i=1}^{n_A} \sum_{k=1,k\neq i}^{n_A} K\left(\frac{u_{A,i}-u_{A,k}}{h}\right) \\ &+ \frac{1}{n_Z(n_Z-1)h} \sum_{i=1}^{n_Z} \sum_{k=1,k\neq i}^{n_Z} K\left(\frac{u_{Z,i}-u_{Z,k}}{h}\right) \\ &- \frac{1}{n_Z(n_A-1)h} \sum_{i=1}^{n_A} \sum_{k=1,k\neq i}^{n_Z} K\left(\frac{u_{A,i}-u_{Z,k}}{h}\right) \\ &- \frac{1}{n_A(n_Z-1)h} \sum_{i=1}^{n_Z} \sum_{k=1,k\neq i}^{n_A} K\left(\frac{u_{Z,i}-u_{A,k}}{h}\right) \end{aligned} \tag{13}$$

$$\begin{aligned} \hat{\sigma}^2_{\gamma,h} =& 2\left\{ \frac{1}{n_A^2 h} \sum_{i=1}^{n_A} \sum_{k=1}^{n_A} K\left(\frac{u_{A,i}-u_{A,k}}{h}\right) \right. \\ &+ \frac{\gamma_n^2}{n_Z^2 h} \sum_{i=1}^{n_Z} \sum_{k=1}^{n_Z} K\left(\frac{u_{Z,i}-u_{Z,k}}{h}\right) \\ &+ \frac{\gamma_n}{n_A n_Z h} \sum_{i=1}^{n_A} \sum_{k=1}^{n_Z} K\left(\frac{u_{A,i}-u_{Z,k}}{h}\right) \\ &\left. + \frac{\gamma_n}{n_Z n_A h} \sum_{i=1}^{n_Z} \sum_{k=1}^{n_A} K\left(\frac{u_{Z,i}-u_{A,k}}{h}\right) \right\} \int K^2(u)du \end{aligned} \tag{14}$$

where: $K(\cdot)$ and h are suitable kernel and suitably selected bandwidth, respectively. $\gamma_n = n_A/n_Z$ and $\lim_{n_A\to\infty} \gamma_n = \gamma, \gamma \in (0,\infty)$.

In addition to the problem of discontinuity discussed above, the main issue of applying Li (1996) test (or any tests based on kernel density estimator or Kolmogorov-Smirnov test) in the context of DEA/FDH is that we are not using the true efficiency scores in the test, but their DEA/FDH estimates, and these estimates are biased and are serially correlated in a complicated unknown way. Simar and Zelenyuk (2006) propose two approaches to remedy the issues. The two approaches are based on bootstrapping the Li (1996) test statistics from: either (i) the sample of estimated efficiency scores with those efficiency scores whose value equal unity being removed (the first approach), or (ii) the sample of the 'smoothed' estimated efficiency scores, i.e., 'smoothing' away from the boundary those efficiency scores whose value equal unity by adding a small uniform noise to them (the second approach) (see the bootstrap algorithms in Simar & Zelenyuk, 2006).

3.2 Analysis of Simple and Weighted Means

To analyse the overall tendency of efficiency in the population, researchers often estimate and perform the statistical inference based on the simple mean of DEA/FDH efficiency scores of a random sample, $\mathcal{S}^n = \{(X^i, Y^i) \mid i = 1, \cdots, n\}$, i.e.,[17]

$$\bar{\theta}_n = \frac{1}{n}\sum_{i=1}^{n} \widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right). \tag{15}$$

This simple estimator for the population mean has a drawback that it does not take the size and thus the economic influence of individual DMUs in the population into account. One way to account for the relative economic importance of individual DMUs in the population is to utilise a weighted mean of DEA/FDH efficiency scores, e.g.,

$$\hat{\tau}_n = \sum_{i=1}^{n} \widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right) \times W^i, \quad W^i = \frac{pY^i}{\sum_{i=1}^{n} pY^i}, \tag{16}$$

where W^i is the weight, and p is a row vector of output prices. The weighted mean formulated above is nothing but an envelopment estimator of the output oriented aggregate technical efficiency proposed by Färe and Zelenyuk (2003),

$$\tau_n = \sum_{i=1}^{n} OTE\left(X^i, Y^i\right) \times W^i, \quad W^i = \frac{pY^i}{\sum_{i=1}^{n} pY^i}. \tag{17}$$

As discussed in Färe and Zelenyuk (2003), the main advantage of this aggregate measure is that the weights are economically meaningful since they are derived from the economic optimisation principle.[18]

[17] We change the notations to the capital letters to highlight that the statistical inference for the population mean needs to be based on the statistical properties of the DEA/FDH estimators of efficiency scores at the random points.

[18] Here we focus our discussion on the analysis of simple and weighted means of DEA/FDH efficiency scores in the whole population. One can find the similar discussion for the case of multiple groups in the population in Nguyen and Zelenyuk (2021a).

It is well-known in the literature that the conventional central limit theorems (CLT) fail to apply to $\bar{\theta}_n$ and $\hat{\tau}_n$ since these estimators inherit the statistical properties of DEA/FDH estimators, which are biased and the bias is of a higher order than the variance when the number of inputs and outputs increases. Kneip et al. (2015) have recently developed the new CLTs for the simple mean of DEA/FDH efficiency scores by correcting the bias and controlling the convergence rates for both bias and variance. Simar and Zelenyuk (2018) extended the important work of Kneip et al. (2015) to the case of the weighted mean of DEA/FDH efficiency scores and developed the new CLTs for this estimator. Simar and Zelenyuk (2020) further improved the finite sample approximation by these new CLTs by proposing to use the bias-corrected estimators of variances. These important works have laid theoretical foundation for many useful statistical inference methods involving DEA/FDH estimators. In this section, we will briefly summarise the important results in these studies focusing on the CLTs and confidence interval estimation.

3.2.1 Central Limit Theorems and Confidence Interval for Simple Means

The first important result developed in Kneip et al. (2015) is the statistical properties of DEA/FDH estimators at a random point, which can be summarised as follows

$$E\left[\widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right) - OTE\left(X^i, Y^i\right)\right] = Cn^{-\kappa} + R_{n,\kappa}, \tag{18}$$

$$E\left[\left(\widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right) - OTE\left(X^i, Y^i\right)\right)^2\right] = o\left(n^{-\kappa}\right), \tag{19}$$

$$\begin{aligned}\Big|COV\Big[&\widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right) - OTE\left(X^i, Y^i\right), \\ &\widehat{OTE}\left(X^j, Y^j|\mathcal{S}^n\right) - OTE\left(X^j, Y^j\right)\Big]\Big| = o\left(n^{-1}\right),\end{aligned} \tag{20}$$

when $n \to \infty$ and under regularity conditions specified in Kneip et al. (2015). The values of the constant term C, the convergence rate κ and the remainder term $R_{n,\kappa}$ are different for different estimators and depend on the dimension of the production space (see Table 1).

With this important result, Kneip et al. (2015) establish the new CLTs for the simple mean of DEA/FDH efficiency scores. Specifically, under regularity conditions specified in Kneip et al. (2015), and for $N + M \leq 5$ if DEA-CRS is used, for $N + M \leq 4$ if DEA-VRS is used, and for $N + M \leq 3$ if FDH is used, the CLTs for the simple mean of DEA/FDH efficiency scores are given by

$$\frac{\sqrt{n}}{\hat{\sigma}_\theta}\left(\bar{\theta}_n - \hat{B}_{\bar{\theta}_n,\kappa} - \mu_\theta + R_{n,\kappa}\right) \xrightarrow{d} N\left(0, 1\right), \tag{21}$$

where $\mu_\theta = E\left[OTE\left(X^i, Y^i\right)\right]$ is the population mean, $\hat{\sigma}_\theta$ is the empirical version of the population standard deviation, $\sigma_\theta = \sqrt{VAR\left[OTE\left(X^i, Y^i\right)\right]}$, and $\hat{B}_{\bar{\theta}_n,\kappa}$ is the generalized jackknife estimator of the bias of $\bar{\theta}_n$, which will be discussed later.

Table 1: Rate of convergence of envelopment estimators

Estimators	κ	$R_{n,\kappa}$
DEA-CRS	$2/(N+M)$	$O\left(n^{-3\kappa/2}(\log n)^{\alpha_1}\right)$
DEA-VRS	$2/(N+M+1)$	$O\left(n^{-3\kappa/2}(\log n)^{\alpha_2}\right)$
FDH	$1/(N+M)$	$O\left(n^{-2\kappa}(\log n)^{\alpha_3}\right)$

The values of $\alpha_j,\ j=1,2,3$, are greater than one and can be found in Kneip et al. (2015)

Alternatively, for $\kappa < 1/2$, the new CLTs are given by

$$\frac{\sqrt{n_\kappa}}{\hat{\sigma}_\theta}\left(\bar{\theta}_{n,\kappa} - \hat{B}_{\bar{\theta}_{n,\kappa}} - \mu_\theta + R_{n,\kappa}\right) \xrightarrow{d} N(0,1) \tag{22}$$

where $\bar{\theta}_{n,\kappa}$ is a subsample version of $\bar{\theta}_n$, in the sense that the averages are taken over a random subsample $\mathcal{S}_{n_\kappa} \subset \mathcal{S}^n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor$. Formally

$$\bar{\theta}_{n,\kappa} = n_\kappa^{-1} \sum_{\{i|(X^i,Y^i)\in\mathcal{S}_{n_\kappa}\}} \widehat{OTE}\left(X^i, Y^i|\mathcal{S}^n\right). \tag{23}$$

The procedure to obtain the generalized jackknife estimator of the bias of $\bar{\theta}_n$ in Kneip et al. (2015) can be summarised as follows. First, let us consider a random split, say random split ℓ, of the sample into two parts with their sizes being $m_1 = \lfloor n/2 \rfloor$ and $m_2 = n - m_1$, and let us denote the random subset of size m_1 of $\mathcal{S}^n$ as $\mathcal{S}^{(1)}_{m_1,\ell}$, and the set of remainders in $\mathcal{S}^n$ as $\mathcal{S}^{(2)}_{m_2,\ell}$.

For $j \in \{1,2\}$, let

$$\bar{\theta}^{(j)}_{m_j,\ell} = (m_j)^{-1} \sum_{(X^i,Y^i)\in\mathcal{S}^{(j)}_{m_j,\ell}} \widehat{OTE}\left(X^i, Y^i|\mathcal{S}^{(j)}_{m_j,\ell}\right). \tag{24}$$

Now, let us define

$$\bar{\theta}^*_{n,\ell} = \frac{1}{2}\left(\bar{\theta}^{(1)}_{m_1,\ell} + \bar{\theta}^{(2)}_{m_2,\ell}\right), \tag{25}$$

then

$$\left(2^\kappa - 1\right)^{-1}\left(\bar{\theta}^*_{n,\ell} - \bar{\theta}_n\right) \tag{26}$$

provides an estimator of the bias term of $\bar{\theta}_n$. To reduce the variance of the estimator, one can repeat the above operations for L times and average (26) to obtain $\hat{B}_{\bar{\theta}_{n,\kappa}}$

$$\hat{B}_{\bar{\theta}_{n,\kappa}} = L^{-1}\sum_{\ell=1}^{L}\left(2^\kappa - 1\right)^{-1}\left(\bar{\theta}^*_{n,\ell} - \bar{\theta}_n\right). \tag{27}$$

Using the new CLTs, one can construct the $(1-\alpha)\,100\%$ confidence interval for the simple mean in the population as

$$\left[\bar{\theta}_n - \hat{B}_{\bar{\theta}_{n,\kappa}} \pm z_{1-\alpha/2}\frac{\hat{\sigma}_\theta}{\sqrt{n}}\right], \tag{28}$$

when (21) can be applied (here $z_{1-\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal distribution), and

$$\left[\bar{\theta}_{n,\kappa} - \hat{B}_{\bar{\theta}_{n,\kappa}} \pm z_{1-\alpha/2}\frac{\hat{\sigma}_\theta}{\sqrt{n_\kappa}}\right], \tag{29}$$

when (22) can be applied.

Moreover, it worth mentioning here that the bias of envelopment estimators is also a source of bias in estimating variance of individual efficiencies. Specifically, Simar and Zelenyuk (2020) show that the empirical version of variance, $\hat{\sigma}^2_\theta$, underestimates the population variance, σ^2_θ. To improve the accuracy of statistical inference in finite sample, Simar and Zelenyuk (2020) propose to use the following bias-corrected version of the estimator of variance

$$\tilde{\sigma}^2_\theta = \hat{\sigma}^2_\theta + \hat{B}^2_{\bar{\theta}_{n,\kappa}}. \tag{30}$$

As discussed in Simar and Zelenyuk (2020), the new CLTs also apply if $\tilde{\sigma}_\theta$ is used in place of $\hat{\sigma}_\theta$ in (21) and (22). Therefore, the $(1-\alpha)\,100\%$ confidence interval for the simple mean in the population can also be obtained by

$$\left[\bar{\theta}_n - \hat{B}_{\bar{\theta}_{n,\kappa}} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}_\theta}{\sqrt{n}}\right], \tag{31}$$

when (21) can be applied, and

$$\left[\bar{\theta}_{n,\kappa} - \hat{B}_{\bar{\theta}_{n,\kappa}} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}_\theta}{\sqrt{n_\kappa}}\right], \tag{32}$$

when (22) can be applied.

3.2.2 Central Limit Theorems and Confidence Interval for Weighted Means

To develop the CLTs for the weighted mean, $\hat{\tau}_n$, Simar and Zelenyuk (2018) first propose an alternative representation of the aggregate efficiency, τ_n, as

$$\tau_n = \frac{(1/n)\sum_{i=1}^n pY^{i\partial}}{(1/n)\sum_{i=1}^n pY^i} = \frac{(1/n)\sum_{i=1}^n Z^{i\partial}}{(1/n)\sum_{i=1}^n Z^i}, \tag{33}$$

where, $Z^i \equiv pY^i$ is the revenue of DMU i, and $Z^{i\partial} \equiv pY^{i\partial} = OTE\left(X^i, Y^i\right) pY^i$ can be viewed as the monetary value of the projection of Y^i on the production frontier. τ_n then can be viewed as an estimator of the population parameter τ, which is a ratio of two population means, i.e.,

$$\tau = \frac{\mu_1}{\mu_2}, \tag{34}$$

where $\mu_1 = E\left[Z^{i\partial}\right]$ and $\mu_2 = E\left[Z^i\right]$. Simar and Zelenyuk (2018) show that the conventional CLT is applied to τ_n. Specifically,

$$\frac{\sqrt{n}}{\sigma_\tau}\left(\tau_n - \tau\right) \xrightarrow{d} N\left(0,1\right), \tag{35}$$

where

$$\sigma_\tau^2 = \tau^2\left(\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2} - 2\frac{\sigma_{12}}{\mu_1\mu_2}\right), \tag{36}$$

where $\sigma_1^2 = VAR\left[Z^{i\partial}\right]$, $\sigma_2^2 = VAR\left[Z^i\right]$, and $\sigma_{12} = COV\left[Z^{i\partial}, Z^i\right]$.

In practice, τ_n is not observable since the true efficiency scores are not observable, thus we need to estimate the population parameter and perform statistical inference about it based on the weighted mean of DEA/FDH efficiency scores, $\hat{\tau}_n$. $\hat{\tau}_n$ can also be rewritten as

$$\hat{\tau}_n = \frac{\hat{\mu}_{1,n}}{\hat{\mu}_{2,n}} = \frac{(1/n)\sum_{i=1}^{n}\hat{Z}^{i\partial}}{(1/n)\sum_{i=1}^{n}Z^i}, \tag{37}$$

where $\hat{Z}^{i\partial} = \widehat{OTE}\left(X^i, Y^i\right)pY^i$.

Simar and Zelenyuk (2018) extend the theory in Kneip et al. (2015) to develop the CLTs for the weighted mean. Specifically, under the regularity conditions specified in Kneip et al. (2015), Theorem 2 in Simar and Zelenyuk (2018) establishes that for $N+M \leq 5$ if DEA-CRS is used, for $N+M \leq 4$ if DEA-VRS is used, and for $N+M \leq 3$ if FDH is used, the CLTs for the weighted mean of DEA/FDH efficiency scores are given by

$$\frac{\sqrt{n}}{\hat{\sigma}_{\tau,n}}\left(\hat{\tau}_n - \hat{B}_{\hat{\tau}_n,\kappa} - \tau + R_{n,\kappa}\right) \xrightarrow{d} N\left(0,1\right) \tag{38}$$

where $\hat{\sigma}_{\tau,n}$ is a consistent estimator of σ_τ given by

$$\hat{\sigma}_{\tau,n}^2 = \hat{\tau}_n\left(\frac{\hat{\sigma}_{1,n}^2}{\hat{\mu}_{1,n}^2} + \frac{\hat{\sigma}_{2,n}^2}{\hat{\mu}_{2,n}^2} - 2\frac{\hat{\sigma}_{12,n}}{\hat{\mu}_{1,n}\hat{\mu}_{2,n}}\right), \tag{39}$$

with $\hat{\sigma}_{1,n}, \hat{\sigma}_{2,n}$, and $\hat{\sigma}_{12,n}$ are empirical versions of σ_1, σ_2, and σ_{12}, respectively. $\hat{B}_{\hat{\tau}_n,\kappa}$ is the generalized jackknife estimator of bias of $\hat{\tau}_n$, which can be obtained using a similar procedure as discussed for the case of the simple mean. Specifically,

$$\hat{B}_{\hat{\tau}_n,\kappa} = \frac{\hat{B}_{\hat{\mu}_{1,n},\kappa}}{\hat{\mu}_{2,n}} = \frac{L^{-1}\sum_{\ell=1}^{L}\left(2^\kappa - 1\right)^{-1}\left(\hat{\mu}_{1,n,\ell}^* - \hat{\mu}_{1,n}\right)}{\hat{\mu}_{2,n}}, \tag{40}$$

where $\hat{\mu}_{1,n,\ell}^*$ is analogous to $\bar{\theta}_{n,\ell}^*$ as defined in (25).

Alternatively, for $\kappa < 1/2$, the CLTs for the weighted mean of DEA/FDH efficiency scores are given by

$$\frac{\sqrt{n_\kappa}}{\hat{\sigma}_{\tau,n}}\left(\hat{\tau}_{n_\kappa} - \hat{B}_{\hat{\tau}_n,\kappa} - \tau + R_{n,\kappa}\right) \xrightarrow{d} N\left(0,1\right), \tag{41}$$

where $\hat{\tau}_{n,\kappa}$ is a subsample version of $\hat{\tau}_n$, in the sense that the averages are taken over a random subsample $\mathcal{S}_{n_\kappa} \subset \mathcal{S}^n$ of size $n_\kappa = \lfloor n^{2\kappa} \rfloor$. Formally

$$\hat{\tau}_{n,\kappa} = \frac{n_\kappa^{-1} \sum_{\{i|(X^i,Y^i)\in\mathcal{S}_{n_\kappa}\}} \hat{Z}^{i\partial}(\mathcal{S}^n)}{n_\kappa^{-1} \sum_{\{i|(X^i,Y^i)\in\mathcal{S}_{n_\kappa}\}} Z^i}, \tag{42}$$

where $\hat{Z}^{i\partial}(\mathcal{S}^n) = \widehat{OTE}(X^i, Y^i|\mathcal{S}^n) Z^i$.

Using the new CLTs, one can construct the $(1-\alpha)\,100\%$ confidence interval for the weighted mean in the population as

$$\left[\hat{\tau}_n - \hat{B}_{\hat{\tau}_n,\kappa} \pm z_{1-\alpha/2}\frac{\hat{\sigma}_{\tau,n}}{\sqrt{n}}\right], \tag{43}$$

when (38) can be applied, and

$$\left[\hat{\tau}_{n,\kappa} - \hat{B}_{\hat{\tau}_n,\kappa} \pm z_{1-\alpha/2}\frac{\hat{\sigma}_{\tau,n}}{\sqrt{n_\kappa}}\right], \tag{44}$$

when (41) can be applied.

As in the case of the simple mean, the bias of envelopment estimators is also a source of bias in estimating the variance of aggregate efficiency. To improve the accuracy of statistical inference in a finite sample, following Simar and Zelenyuk (2020), one can use the bias-corrected version of the estimator of variance

$$\tilde{\sigma}^2_{\tau,n} = \hat{\tau}_n \left(\frac{\tilde{\sigma}^2_{1,n}}{\hat{\mu}^2_{1,n}} + \frac{\hat{\sigma}^2_{2,n}}{\hat{\mu}^2_{2,n}} - 2\frac{\hat{\sigma}_{12,n}}{\hat{\mu}_{1,n}\hat{\mu}_{2,n}}\right), \tag{45}$$

where

$$\tilde{\sigma}^2_{1,n} = \hat{\sigma}^2_{1,n} + \hat{B}^2_{\hat{\mu}_{1,n},\kappa}. \tag{46}$$

The new CLTs also apply if $\tilde{\sigma}_{\tau,n}$ is used in place of $\hat{\sigma}_{\tau,n}$ in (38) and (41). Therefore, the $(1-\alpha)\,100\%$ confidence interval for the weighted mean in the population can also be obtained by

$$\left[\hat{\tau}_n - \hat{B}_{\hat{\tau}_n,\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}_{\tau,n}}{\sqrt{n}}\right], \tag{47}$$

when (38) can be applied, and

$$\left[\hat{\tau}_{n,\kappa} - \hat{B}_{\hat{\tau}_n,\kappa} \pm z_{1-\alpha/2}\frac{\tilde{\sigma}_{\tau,n}}{\sqrt{n_\kappa}}\right], \tag{48}$$

when (41) can be applied.

As demonstrated by intensive Monte Carlo simulations in Simar and Zelenyuk (2020), for both the case of simple mean and weighted mean, the confidence intervals constructed using bias-corrected version of variance estimator have a better coverage and thus are more reliable.

4 Data and Variables

4.1 Sample

For this study we were able to obtain a fairly disaggregated dataset on public hospitals in Queensland that, among other things, consists of 153,472 data points for 704 different types of inpatient procedures across 109 hospitals in two years. To our understanding of the literature, it is a relatively rich dataset as researchers often have to use a much more aggregated data such as patient days (e.g., Sherman, 1984; Bates et al., 2006; Hu et al., 2012) or the number of patients or admissions or discharges (e.g., Hao & Pegels, 1994; Zuckerman et al., 1994; Athanassopoulos et al., 1999; McCallion et al., 1999; Grosskopf et al., 2001; Staat, 2006; Tiemann & Schreyögg, 2009; Besstremyannaya, 2013). Indeed, despite many attempts we also were not able to obtain similar level of data for other States or Territories of Australia. So, while we cannot generalise our conclusions for other States, Territories or especially other countries, our results for Queensland might still be insightful, at least as a food for thought, for researchers across the globe.

Specifically, the sample for our analysis consists of 109 public acute hospitals in Queensland, Australia in the two financial years (FY) 2012/13 and 2013/2014.[19] The data is provided by Queensland Department of Health, extracted from the two data collections which are the Financial and Residential Activity Collection (FRAC) and the Monthly Activity Collection (MAC).

4.2 Variables

The list of variables used in this study is shown in Table 2 (with their notations and descriptions). In the following subsections, we will discuss each variable in details.

4.2.1 Inpatient Output

The main variable of interest in this analysis is the inpatient output. In our data, inpatient output is recorded in terms of admitted patient episodes, which are categorised into 704 Diagnosis-Related Groups (DRGs).

In principle, each DRG represents a separate type of output, yet they need to be aggregated to make the model feasible. Here, we consider two approaches to aggregate the inpatient output: (i) PCA-based approach and (ii) Price-based approach. Moreover, for the price-based approach, we consider three different weight systems, which are: (i) varying DRG cost weights (different weights for different years), (ii) constant cost weights

[19] In Australia, a financial year starts on 1 July and ends on 30 June of the next calendar year.

Table 2: The description of variables

Variables	Description
Inputs	
FLABOUR	Labour input (PCA-based aggregation)
BEDS	Capital input (Total beds)
DMSUP	Consumable input (Drug and medical supply expenditure) (2012/13 constant price)($1,000,000s)
Outputs	
OUT	Outpatient output (Non-admitted occasions of service)(1,000s)
WEPS1	Inpatient output (Price-based aggregation using varying weights) (1,000s)
WEPS2	Inpatient output (Price-based aggregation using 2012/13 weights) (1,000s)
WEPS3	Inpatient output (Price-based aggregation using 2013/14 weights) (1,000s)
FEPS	Inpatient output (PCA-based aggregation)
TEPS	Inpatient output (Non-weighted aggregation)(1,000s)

using FY 2012/13 cost weights, and (iii) constant cost weights using FY 2013/14 cost weights.[20] Besides these, we also compare the results from PCA-based and price-based approaches with those from a non-weighted aggregation approach, in which the inpatient output is measured by the raw count of the total number of episodes.

The relationship between the cost weights in FY 2012/13 and FY 2013/14 is shown in Figure 1. We can see that there is a variation (but small in magnitude) in the cost weights across the years, and thus it is useful to assess the sensitivity of the results with respect to the different weight systems as was done in the simulation exercises in Zelenyuk (2020).

Figure 2 shows the total number of episodes of each DRG and the number of hospitals producing the DRG in 2013/14.[21] It can be seen from the figure that DRGs with high weights are produced in a small quantity and at few hospitals, whereas the DRGs that are produced in a massive quantity at many hospitals usually have low weights. The DRGs with high cost weights typically include highly specialised surgical procedures,

[20] It is worth mentioning here that under the activity-based funding system in Australia, each inpatient DRG has a single cost weight, which is updated annually based on historical costs incurred by hospitals in treating patients belonging to the DRG. An inpatient episode in each DRG is then funded at a flat rate based upon the DRG cost weight and the so-called national efficient price (NEP)–a fixed price paid per weighted unit.

[21] The patterns in the year 2012/13 are similar, to save space, we only present a graph for the year 2013/14

such as Insertion of Ventricular Assist Device (A10Z), Liver Transplant (A01Z), and so on. Meanwhile, the ones with low cost weights are usually simple medical procedures, such as Electroconvulsive Therapy (U60Z), Haemodialysis (L61Z), and so on.

With regard to the PCA-based approach, the aggregate inpatient output (i.e., the first factorial variable) contains 65.39% information in all 704 DRGs as $\delta_1 = 0.6539$. Interestingly, in spite of the fact that the weights from the PCA-based approach are negatively correlated with the DRG cost weights [22], there are almost perfect positive correlations among the four relevant aggregated outputs (i.e., WEPS1, WEPS2, WEPS2 and FEPS), as well as between these four aggregated outputs and the non-weighted aggregated output (TEPS) (e.g., see the pairwise Pearson linear correlation coefficients in Table 3). The relationships among the aggregated outputs discussed here might make one conjecture that the nature of the dataset makes the results of analysis indifferent across all aggregation approaches regardless whether they are weighted or non-weighted, but it is not necessarily the case and we will examine this conjecture in Section 5.

Table 3: Pairwise Pearson linear correlation coefficients among five aggregated inpatient outputs

	WEPS1	WEPS2	WEPS3	FEPS	TEPS
WEPS1	1				
WEPS2	0.9997	1			
WEPS3	0.9997	0.9995	1		
FEPS	0.9966	0.9958	0.9974	1	
TEPS	0.9813	0.9800	0.9828	0.9867	1

[22] E.g., pairwise Pearson linear correlation coefficients between PCA-based weights and DRG cost weights are around -0.33.

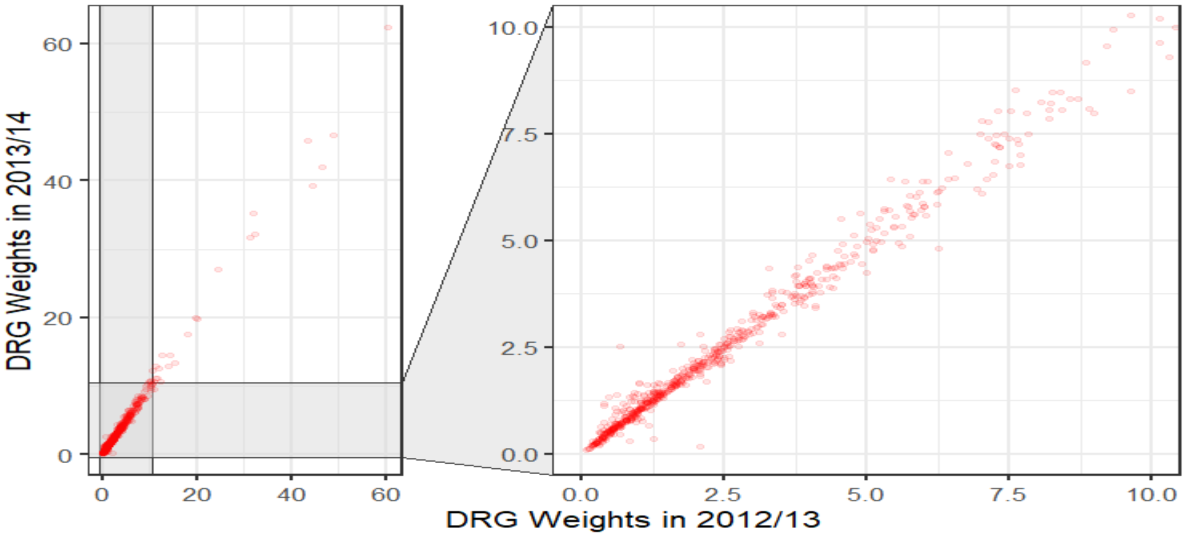

Figure 1: Scatter plot of DRG weights in 2012/13 vs. 2013/14. The right panel is a zoomed part of the left panel.

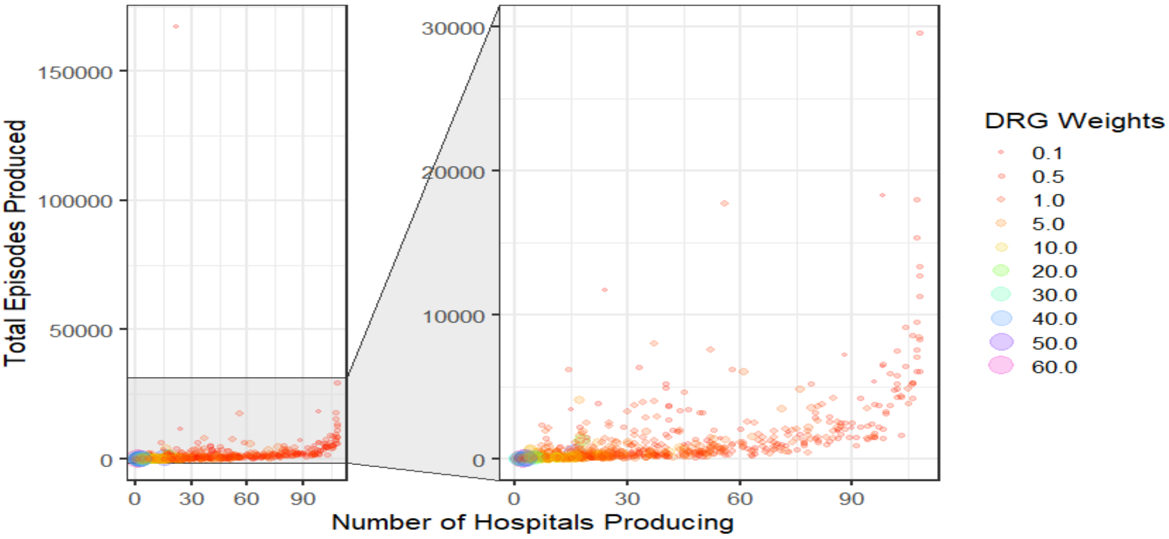

Figure 2: Total number of episodes of each DRG and the number of hospitals producing the DRG in 2013/14. Each point reprents a DRG. The colour and size of the points represent DRGs' weights. The right panel is a zoomed part of the left panel.

4.2.2 Outpatient Output

Another hospital output used in this study is outpatient output (OUT). Hospital outpatient output is measured by a raw count of the number of non-admitted occasions of service including both emergency and non-emergency services for non-admitted patients. The outpatient output also shows a strong positive correlation with each of the five aggregated inpatient outputs with the pairwise Pearson linear correlation coefficients being around 0.97 to 0.99.

4.2.3 Inputs

We use three inputs to model the production process of hospitals, which are: labour input (PCA-based aggregation of full time equivalent staff in six labour categories), capital input (proxied by the number of beds), and consumable input (proxied by drug, surgical and medical supply expenditure). On the input side, this is therefore a mixture of PCA-based approach and price-based approach.

The descriptive statistics of variables used in the study are shown in Table 4. As in any typical healthcare related data, we observe great variation and positive skewness for all variables in the working sample.

Table 4: The descriptive statistics of variables

Variables	Mean	Median	Q1	Q3	Std. Dev.	Min	Max
Inputs							
FLABOUR	1.01	0.12	0.07	0.42	2.43	0.02	15.56
BEDS	92.55	20.00	10.00	48.00	185.42	5.00	1024.00
DMSUP	10.12	0.42	0.15	2.37	27.45	0.03	167.35
Outputs							
OUT	88.60	19.88	6.31	52.23	166.02	1.31	954.90
WEPS1	9.93	0.87	0.31	4.05	23.05	0.03	132.61
WEPS2	9.94	0.83	0.29	3.93	23.18	0.03	132.61
WEPS3	9.90	0.88	0.31	4.05	22.87	0.03	129.36
FEPS	9.35	1.04	0.43	3.94	20.84	0.043	115.25
TEPS	9.07	1.02	0.38	4.58	18.37	0.05	95.58

5 A comparison of estimated efficiency scores among different aggregation approaches

In the following analysis, we compare the DEA-CRS estimates of hospital efficiency across the five models in which three inputs and two outputs are utilised (see model specifications in Table 5).[23]

Table 5: Model Specifications for the case of 3-inputs and 2-outputs models

Model	Inputs			Outputs					
	FLABOUR	BEDS	DMSUP	OUT	WEPS1	WEPS2	WEPS3	FEPS	TEPS
Model 1	✓	✓	✓	✓	✓				
Model 2	✓	✓	✓	✓		✓			
Model 3	✓	✓	✓	✓			✓		
Model 4	✓	✓	✓	✓				✓	
Model 5	✓	✓	✓	✓					✓

Table 6 shows the descriptive statistics on the central tendency and dispersion of the estimated efficiency scores from the five models. At first glance, we can see that the descriptive statistics for the estimated efficiency scores from Model 1, Model 2, and Model 3, where the inpatient outputs are aggregated using price-based approaches with different weight systems, are almost identical. To further investigate the relationship among the estimated efficiency scores of these three models, we look at their pairwise Spearman rank-order correlation coefficients. It is shown in Table 7 that the estimated efficiency scores are nearly perfectly correlated with all the correlation coefficients being around 0.97-0.99, implying that the rankings of the individual hospitals based on the estimated efficiency scores from the three models are almost the same.

We then compare the 95% confidence intervals for the simple and weighted means of efficiency scores estimated from these three models. Table 8 reports the results for

[23] As a sensitivity analysis, we also consider the case of 3-inputs and 1-output models, where we aggregate outpatient output and inpatient output using PCA-based aggregation, and the case of 1-input and 1-output models, where we further aggregate the three inputs into a single measure of hospital inputs using PCA-based aggregation. The results from the sensitivity analysis are qualitatively similar to the discussion in this section (see the results in the Appendix).

the simple mean and Table 9 presents the estimates for the weighted means.[24] We can see that for both the simple mean and the weighted mean, the biased corrected DEA estimates as well as the 95% confidence intervals are almost the same between Model 1 and Model 2. Compared to the first two models, the results from Model 3 are slightly different for the case of simple mean, and the difference is more pronounced when we consider the weighted mean. Specifically, the 95% confidence interval of the weighted mean from Model 3 is to the left and does not overlap with those from Model 1 and Model 2. However, the difference here might just be particular for the dataset used in this study. For this dataset, when using the fixed weights of FY2013/14, the quantity of the aggregate output of most (but not all) observations in FY2012/13 increases, but the estimated production frontier is moving inward due to a particular observation. This observation is on the estimated frontiers no matter whether WEPS1, WEPS2, or WEPS3 is used in the DEA models, and its value of WEPS3 is significantly smaller than WEPS1 and WEPS2. As a result, the estimated frontier of Model 3 around this observation is significantly lower than those of Model 1 and Model 2. Consequently, the estimated efficiency scores of other observations in this segment of the production space are smaller in Model 3 compared to Model 1 and Model 2. Moreover, the observation that we are referring to is relatively large (in terms of revenue), and thus so do its peers, and this is the reason why the difference is more pronounced when we compare the weighted mean. This is also a good real data example of the importance of accounting for the economic weight of each efficiency score in the averaging of the scores over observations of very different sizes.

[24] It is worth mentioning here that to perform the analysis for the weighted mean, it requires information about the relative price between the outpatient output and the aggregate inpatient output. For the price-based aggregate inpatient output, we use the information about the relative price in Nguyen and Zelenyuk (2021a). However, the price information is not available for PCA-based aggregate inpatient output as well as for the raw count of inpatient episodes, thus we focus only on the simple mean for Model 4 and Model 5.

Table 6: Descriptive statistics of estimated efficiency scores using DEA-CRS estimator for the case of 3-inputs and 2-outputs models

Model	Mean	Median	Q1	Q3	Std. Dev.	Min	Max
Model 1	1.55	1.38	1.22	1.70	0.54	1.00	4.08
Model 2	1.54	1.37	1.21	1.70	0.53	1.00	4.05
Model 3	1.52	1.34	1.17	1.71	0.54	1.00	4.08
Model 4	1.61	1.43	1.23	1.80	0.58	1.00	4.68
Model 5	1.73	1.58	1.24	1.97	0.70	1.00	5.15

Table 7: Pairwise Spearman rank-order correlation coefficients of estimated efficiency scores using DEA-CRS estimator across models for the case of 3-inputs and 2-outputs models

	Model 1	Model 2	Model 3	Model 4	Model 5
Model 1	1.00				
Model 2	0.99	1.00			
Model 3	0.97	0.98	1.00		
Model 4	0.94	0.94	0.93	1.00	
Model 5	0.89	0.90	0.87	0.90	1.00

Table 8: 95% confidence intervals for the simple mean of DEA-CRS efficiency scores for the case of 3-inputs and 2-outputs models

	Model 1	Model 2	Model 3	Model 4	Model 5
DEA estimate	1.55	1.54	1.52	1.61	1.73
Bias corrected	1.83	1.81	1.75	1.90	2.04
Std. Error	0.06	0.06	0.06	0.07	0.08
Bias corrected Std. Error	0.07	0.07	0.07	0.08	0.09
LB of est. CI	1.69	1.68	1.62	1.73	1.84
UB of est. CI	1.94	1.92	1.87	2.00	2.15
LB of est. CI-Improved	1.68	1.67	1.61	1.72	1.82
UB of est. CI-Improved	1.95	1.94	1.88	2.02	2.17

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Table 9: 95% confidence intervals for the weighted mean of DEA-CRS efficiency scores for the case of 3-inputs and 2-outputs models

	Model 1	Model 2	Model 3	Model 4	Model 5
DEA estimate	1.31	1.30	1.19	n/a	n/a
Bias corrected	1.58	1.55	1.28	n/a	n/a
Std. Error	0.03	0.03	0.03	n/a	n/a
Bias corrected Std. Error	0.04	0.04	0.03	n/a	n/a
LB of est. CI	1.51	1.48	1.23	n/a	n/a
UB of est. CI	1.64	1.61	1.34	n/a	n/a
LB of est. CI-Improved	1.49	1.46	1.23	n/a	n/a
UB of est. CI-Improved	1.66	1.63	1.34	n/a	n/a

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Furthermore, we examine the whole distribution of estimated efficiency scores from the three models by comparing their estimated densities and performing the adapted Li test for the equality of the densities. From Figure 3, we see that the estimated densities of estimated efficiency scores from Model 1 and Model 2 are almost overlapped. And, although the estimated densities from Model 3 is slightly to the left of those from Model 1 and Model 2, the adapted Li test shows that there is insufficient statistical evidence to conclude that the densities of the estimated efficiency scores from these three models are pairwise different. All in all, the analysis suggests that the efficiency scores estimated from the three models, in which inpatient outputs are aggregated using different weight systems, provide the same information about the overall efficiency level as well as the relative rankings of individual hospitals in our sample. This implies the robustness of results with respect to a slight change in relative prices.

Table 10: The adapted Li test for equality of distributions of DEA-CRS efficiency scores by different models for the case of 3-inputs and 2-outputs models

	The Li test statistics	Bootstrap p-value	Decision (at 5% level of significance)
Model 1 vs. Model 2	-0.01	0.99	Do not reject H_0
Model 1 vs. Model 3	1.30	0.06	Do not reject H_0
Model 2 vs. Model 3	0.96	0.11	Do not reject H_0
Model 1 vs. Model 4	0.04	0.97	Do not reject H_0
Model 2 vs. Model 4	0.19	0.79	Do not reject H_0
Model 3 vs. Model 4	1.76	0.03	Reject H_0
Model 1 vs. Model 5	4.47	0.00	Reject H_0
Model 2 vs. Model 5	4.63	0.00	Reject H_0
Model 3 vs. Model 5	5.21	0.00	Reject H_0
Model 4 vs. Model 5	2.69	0.01	Reject H_0

Notes:

* H_0: The densities of estimated efficiency scores of the two corresponding models are equal.

* Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2006), with 2000 bootstrap replications using Gaussian kernel, and Silverman (1986) robust rule of thumb bandwidth.

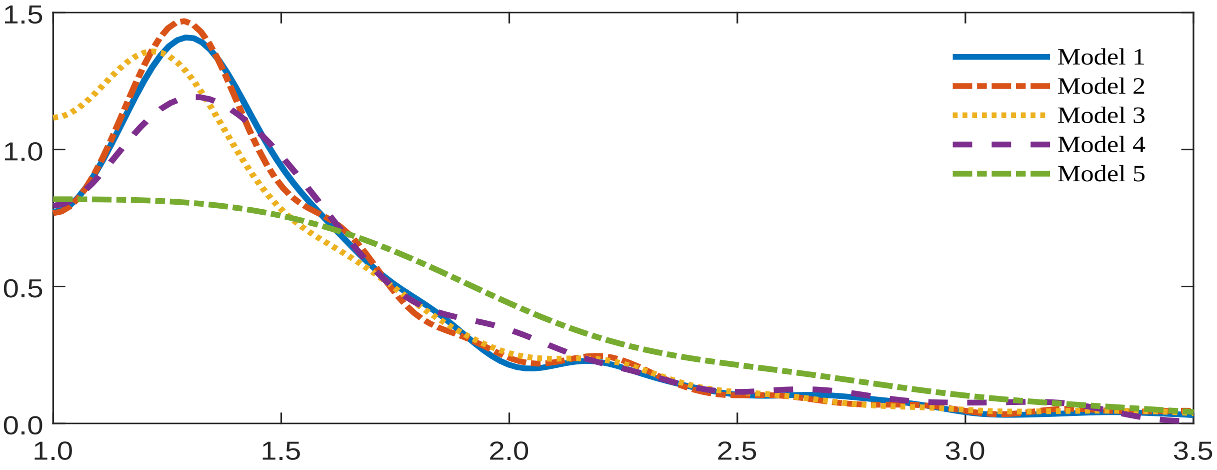

Figure 3: Density estimates of the estimated efficiency scores using the DEA-CRS estimators for different models for the case of 3-inputs and 2-outputs models. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

Let us now look at the estimated efficiency scores from Model 4, where inpatient outputs are aggregated using the PCA-based approach and let us compare the estimates with those from Model 1, Model 2, and Model 3. We can see that although the simple mean of estimated efficiency scores from Model 4 are slightly higher (i.e., indicating overall less efficient) than those from the first three models, the 95% confidence intervals for the population mean constructed using the estimated efficiency scores from the four models almost overlap. Besides, the pairwise Spearman rank-order correlation coefficients between the estimates from Model 4 and each of those from the first three models are still very high (all around 0.93-0.94). Moreover, the estimated density of the estimated efficiency scores from Model 4 looks similar to those from the first three models, especially from Model 1 and Model 2. The adapted Li test supports the visualisation, showing no statistical pairwise difference between the densities of the estimated efficiency scores from Model 1, Model 2 and Model 4. The results suggest that the estimated efficiency scores from the DEA models are also robust with respect to the two different aggregation approaches.

To examine the conjecture discussed in Section 4 that due to the nature of the data, the results might be indifferent regardless of whether the weighted or non-weighted aggregation approaches are utilised, we compare the estimated efficiency scores from Model 5 with those from the first four models, focusing on their whole distribution. We can see from Figure 3 that the estimated density of the efficiency scores obtained from Model 5 is very different from the estimated densities of those obtained from the first four models. Our visualisation is then confirmed by the adapted Li test, where the null hypotheses about the equality between densities of estimated efficiency scores from Model 5 and from each of the first four models are all rejected with very small p-values. The results highlight the importance of utilising the appropriate approaches for dimension reduction in the DEA context: estimated efficiency scores under PCA-based aggregation and price-based aggregation approaches are similar, but very different compared with those obtained under the naive non-weighted approach.

6 Concluding Remarks

Given the increasing attention to dimension reduction in the context of DEA with large dimensions for inputs and outputs, in this study we revisit the output aggregation in hospital efficiency analysis, where the challenge of big wide data for DEA has been present since the early 1980s. Although not cast explicitly, researchers in the field usually use price-based aggregation to aggregate a hospital's output to deal with the most challenging issue of DEA with big data – the 'curse of dimensionality'.

Using data on public hospitals in Queensland, Australia, we find that the choice of price systems (with small variation in prices) does not significantly affect the DEA estimates under the price-based aggregation approach. Moreover, the estimated efficiency scores from the DEA models are also robust with respect to the price-based aggregation and the PCA-based aggregation approaches. The robustness of the results suggests that the PCA-based aggregation can be viewed as a viable alternative for DEA practitioners who are unable to/or unwilling to use the price-based approach, e.g., due to unavailable or unreliable price information.

It is worth highlighting here that the results in this study are particular for the working sample, and should be interpreted with care. Even though, this paper can be viewed as a pilot study to remind DEA practitioners, especially those who work with hospital data, that price-based aggregation is just one approach (among others) to reduce the dimension of production space to deal with the 'curse of dimensionality'. Thus in applied works, several approaches should be applied to ensure the robustness of results.

Recently, the Least Absolute Shrinkage and Selection Operator (LASSO) has been applied into the DEA context as a promising technique for dimension reduction (see more discussions in Chen et al., 2020; Lee & Cai, 2020, and references therein). As a result, a fruitful direction for future research is to empirically compare the DEA estimates based on the LASSO approach with those based on price-based and PCA-based aggregation approaches.

Acknowledgments

We thank the Editors/referees for fruitful comments. We acknowledge the support from The University of Queensland and the financial support from the Australian Research Council (FT170100401). We also thank Zhichao Wang and Evelyn Smart for their valuable feedback. We acknowledge and thank Queensland Health for providing part of the data that we used in this study. These individuals and organizations are not responsible for the views expressed in this paper.

Appendix

A Results for the case of 3-inputs and 1-output models

Here we aggregate outputs into a single measure of hospital outputs using the PCA-based approach and estimate 3-inputs and 1-output production models using DEA-CRS. Specifically, we aggregate the outpatient output with each of the four aggregated inpatient outputs (i.e., WEPS1, WEPS2, WEPS3, and FEPS) and denote the aggregated outputs as FOUT1, FOUT2, FOUT3, and FOUT4, respectively (see model specifications in Table 11)

Table 11: Model Specifications for the case of 3-inputs and 1-output models

Model	Inputs			Outputs				
	FLABOUR	BEDS	DMSUP	FOUT1	FOUT2	FOUT3	FOUT4	FOUT5
Model 6	✓	✓	✓	✓				
Model 7	✓	✓	✓		✓			
Model 8	✓	✓	✓			✓		
Model 9	✓	✓	✓				✓	
Model 10	✓	✓	✓					✓

Table 12: Descriptive statistics of estimated efficiency scores using DEA-CRS estimator for the case of 3-inputs and 1-output models

Model	Mean	Median	Q1	Q3	Std. Dev	Min	Max
Model 6	2.61	2.49	1.98	3.07	1.01	1.00	8.43
Model 7	2.63	2.50	1.99	3.07	1.02	1.00	8.48
Model 8	2.60	2.49	1.95	3.07	1.00	1.00	8.41
Model 9	2.49	2.41	1.87	2.87	0.92	1.00	7.71
Model 10	2.36	2.22	1.74	2.81	0.93	1.00	8.10

Table 13: 95% confidence intervals for the simple mean of DEA-CRS efficiency scores for the case of 3-inputs and 1-outputs model

	Model 6	Model 7	Model 8	Model 9	Model 10
DEA estimate	2.61	2.63	2.60	2.49	2.36
Bias corrected	3.43	3.46	3.41	3.26	3.00
Std. Error	0.07	0.07	0.07	0.06	0.06
Bias corrected Std. Error	0.09	0.09	0.09	0.08	0.08
LB of est. CI	3.30	3.32	3.27	3.13	2.88
UB of est. CI	3.57	3.59	3.54	3.38	3.12
LB of est. CI-Improved	3.26	3.28	3.24	3.10	2.85
UB of est. CI-Improved	3.60	3.63	3.58	3.42	3.15

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Table 14: 95% confidence intervals for the weighted mean of DEA-CRS efficiency scores for the case of 3-inputs and 1-output models

	Model 6	Model 7	Model 8	Model 9	Model 10
DEA estimate	2.66	2.66	2.65	2.65	2.54
Bias corrected	3.73	3.72	3.71	3.66	3.49
Std. Error	0.05	0.05	0.05	0.06	0.07
Bias corrected Std. Error	0.09	0.09	0.09	0.09	0.10
LB of est. CI	3.63	3.62	3.61	3.55	3.34
UB of est. CI	3.83	3.82	3.81	3.77	3.63
LB of est. CI-Improved	3.55	3.55	3.54	3.48	3.29
UB of est. CI-Improved	3.90	3.90	3.88	3.83	3.68

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Table 15: Pairwise Spearman rank-order correlation coefficients of the estimated efficiency scores using the DEA-CRS estimator across models for the case of 3-inputs and 1-output models

	Model 6	Model 7	Model 8	Model 9	Model 10
Model 6	1				
Model 7	0.999	1			
Model 8	0.999	0.999	1		
Model 9	0.982	0.979	0.983	1	
Model 10	0.975	0.973	0.977	0.979	1

Table 16: The adapted Li test for the equality of distributions of DEA-CRS efficiency scores by different models for the case of 3-inputs and 1-output models

	The Li test statistics	Bootstrap p-value	Decision (at 5% level of significance)
Model 6 vs. Model 7	0.006	0.994	Do not reject H_0
Model 6 vs. Model 8	-0.001	0.998	Do not reject H_0
Model 7 vs. Model 8	0.022	0.974	Do not reject H_0
Model 6 vs. Model 9	0.417	0.573	Do not reject H_0
Model 7 vs. Model 9	0.404	0.595	Do not reject H_0
Model 8 vs. Model 9	0.378	0.620	Do not reject H_0
Model 6 vs. Model 10	2.131	0.016	Reject H_0
Model 7 vs. Model 10	2.221	0.013	Reject H_0
Model 8 vs. Model 10	2.012	0.018	Reject H_0
Model 9 vs. Model 10	1.229	0.071	Do not reject H_0

Notes:

* H_0: The densities of estimated efficiency scores of the two corresponding models are equal.

* Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2006), with 2000 bootstrap replications using Gaussian kernel, and Silverman (1986) robust rule of thumb bandwidth.

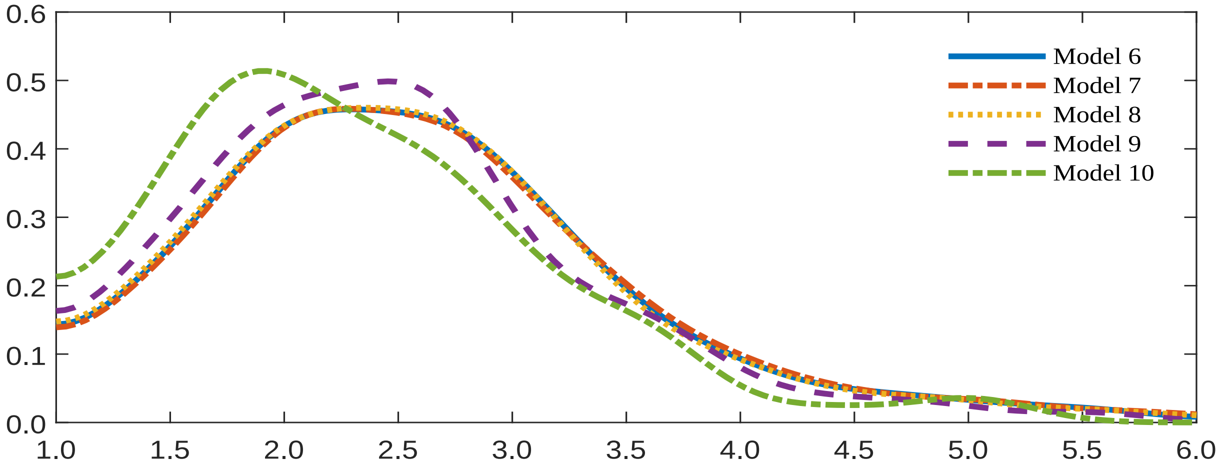

Figure 4: Density estimates of the estimated efficiency scores using the DEA-CRS estimators for different models for the case of 3-inputs and 1-output models. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

B Results for the case of 1-input and 1-output models

Here we further aggregate three inputs into a single measure of input using the PCA-based approach, denoted as FINPUT, and estimate 1-input and 1-output production models using DEA-CRS (see model specifications in Table 17)

Table 17: Model Specifications for the case of 1-input and 1-output models

Model	Inputs	Outputs				
	FINPUT	FOUT1	FOUT2	FOUT3	FOUT4	FOUT5
Model 11	✓	✓				
Model 12	✓		✓			
Model 13	✓			✓		
Model 14	✓				✓	
Model 15	✓					✓

Table 18: Descriptive statistics of the estimated efficiency scores using the DEA-CRS estimator for the case of 1-input and 1-output models

Model	Mean	Median	Q1	Q3	Std. Dev	Min	Max
Model 11	4.82	4.09	3.21	5.46	2.37	1.00	14.23
Model 12	4.86	4.10	3.23	5.47	2.41	1.00	14.23
Model 13	4.78	4.05	3.19	5.42	2.33	1.00	13.75
Model 14	4.46	3.85	3.07	5.27	2.06	1.00	12.21
Model 15	4.28	3.69	2.79	5.14	2.11	1.00	12.76

Table 19: 95% confidence intervals for the simple mean of DEA-CRS efficiency scores for the case of 1-input and 1-outputs model

	Model 11	Model 12	Model 13	Model 14	Model 15
DEA estimate	4.82	4.86	4.78	4.46	4.28
Bias corrected	5.61	5.64	5.55	5.15	4.94
Std. Error	0.16	0.16	0.16	0.14	0.14
Bias corrected Std. Error	0.17	0.17	0.17	0.15	0.15
LB of est. CI	5.29	5.32	5.24	4.88	4.66
UB of est. CI	5.92	5.96	5.86	5.42	5.22
LB of est. CI-Improved	5.28	5.31	5.22	4.86	4.65
UB of est. CI-Improved	5.94	5.98	5.87	5.44	5.24

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Table 20: 95% confidence intervals for the weighted mean of DEA-CRS efficiency scores for the case of 1-input and 1-output models

	Model 11	Model 12	Model 13	Model 14	Model 15
DEA estimate	4.18	4.18	4.17	4.15	3.99
Bias corrected	4.86	4.86	4.84	4.79	4.60
Std. Error	0.13	0.13	0.13	0.14	0.17
Bias corrected Std. Error	0.14	0.14	0.14	0.15	0.18
LB of est. CI	4.60	4.60	4.59	4.51	4.26
UB of est. CI	5.11	5.11	5.10	5.06	4.95
LB of est. CI-Improved	4.59	4.59	4.57	4.50	4.25
UB of est. CI-Improved	5.13	5.12	5.12	5.08	4.96

*LB: Lower Bound, UB: Upper Bound, Est.: Estimated, Std. Error: Standard Error, CI: Confidence Interval.

*Computations are done in the Matlab adopting the code from Simar and Zelenyuk (2020).

Table 21: Pairwise Spearman rank-order correlation coefficients of the estimated efficiency scores using the DEA-CRS estimator across models for the case of 1-input and 1-output models

	Model 11	Model 12	Model 13	Model 14	Model 15
Model 11	1				
Model 12	0.999	1			
Model 13	0.999	0.999	1		
Model 14	0.988	0.987	0.990	1	
Model 15	0.981	0.980	0.982	0.985	1

Table 22: The adapted Li test for the equality of distributions of DEA-CRS efficiency scores by different models for the case of 1-input and 1-output models

	The Li test statistics	Bootstrap p-value	Decision (at 5% level of significance)
Model 11 vs. Model 12	-0.004	0.995	Do not reject H_0
Model 11 vs. Model 13	0.001	0.998	Do not reject H_0
Model 12 vs. Model 13	0.009	0.989	Do not reject H_0
Model 11 vs. Model 14	0.372	0.577	Do not reject H_0
Model 12 vs. Model 14	0.426	0.518	Do not reject H_0
Model 13 vs. Model 14	0.271	0.700	Do not reject H_0
Model 11 vs. Model 15	3.103	0.004	Reject H_0
Model 12 vs. Model 15	3.276	0.002	Reject H_0
Model 13 vs. Model 15	2.864	0.008	Reject H_0
Model 14 vs. Model 15	1.419	0.050	Reject H_0

Notes:

* H_0: The densities of the estimated efficiency scores of the two corresponding models are equal.

* Computations are done in Matlab adopting the code from Simar and Zelenyuk (2006), with 2000 bootstrap replications using Gaussian kernel, and Silverman (1986) robust rule of thumb bandwidth.

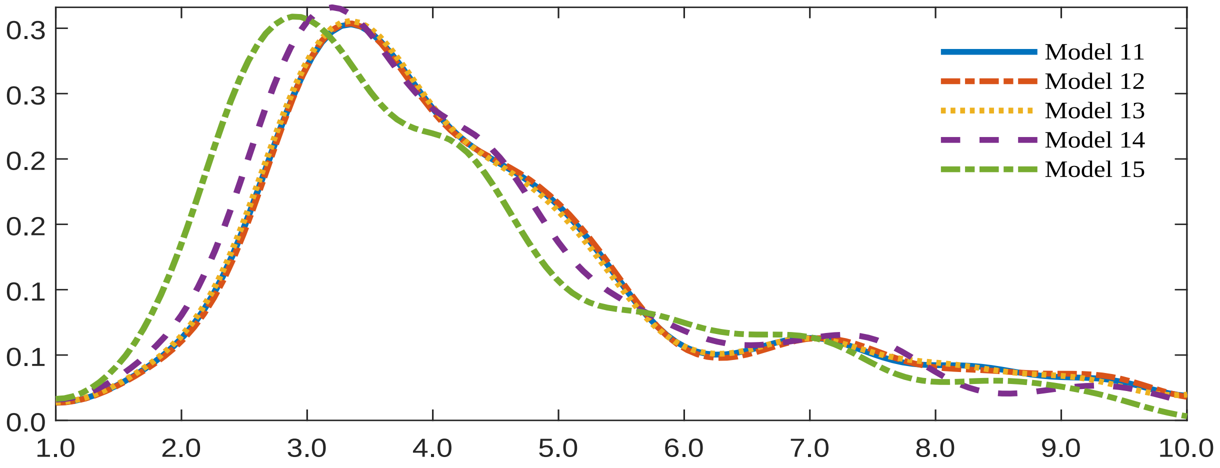

Figure 5: Density estimates of the estimated efficiency scores using the DEA-CRS estimators for different models for the case of 1-input and 1-output models. Kernel-based with Silverman's (1986) reflection method: Gaussian kernel and bandwidth is selected by the method of Sheather and Jones (1991).

References

Adler, N., & 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.

Adler, N., & Golany, B. (2007). PCA-DEA: Reducing the curse of dimensionality. In J. Zhu & W. D. Cook (Eds.), *Modeling data irregularities and structural complexities in data envelopment analysis* (pp. 139–153). Springer US.

Aigner, D., Lovell, C. A. K., & Schmidt, P. (1977). Formulation and astimation of stochastic frontier production function models. *Journal of Econometrics*, *64*(6), 1263–1297.

Aragon, Y., Daouia, A., & Thomas-Agnan, C. (2005). Nonparametric frontier estimation: A conditional quantile-based approach. *Econometric Theory*, *21*(2), 358–389.

Athanassopoulos, A., Gounaris, C., & Sissouras, A. (1999). A descriptive assessment of the production and cost efficiency of general hospitals in Greece. *Health Care Management Science*, *2*(2), 97–106.

Bates, L. J., Mukherjee, K., & Santerre, R. E. (2006). Market structure and technical efficiency in the hospital services industry: A DEA approach. *Medical Care Research and Review*, *63*(4), 499–524.

Besstremyannaya, G. (2013). The impact of Japanese hospital financing reform on hospital efficiency: A difference-in-difference approach. *The Japanese Economic Review*, *64*(3), 337–362.

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

Charles, V., Aparicio, J., & Zhu, J. (2019). The curse of dimensionality of decision-making units: A simple approach to increase the discriminatory power of data envelopment analysis. *European Journal of Operational Research*, *279*(3), 929–940.

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

Chen, Y., Tsionas, M., & Zelenyuk, V. (2020). LASSO DEA for small and big data. *CEPA Working Paper Series*, (WP02/2020).

Daouia, A., & Simar, L. (2007). Nonparametric efficiency analysis: A multivariate conditional quantile approach. *Journal of Econometrics*, *140*(2), 375–400.

Daraio, C., & Simar, L. (2007). Economies of scale, scope and experience in the italian motor-vehicle sector. In C. Daraio & L. Simar (Eds.), *Advanced robust and non-*

parametric methods in efficiency analysis: Methodology and applications (pp. 135–165). Springer Science & Business Media.

Deprins, D., Simar, L., & Tulkens, H. (1984). Measuring labor efficiency in post offices. In M. G. Marchand, P. Pestieau, & H. Tulkens (Eds.), *The performance of public enterprises: Concepts and measurements* (pp. 243–267). Amsterdam, North-Holland.

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

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

Färe, R., & Primont, D. (1995). *Multi-output production and duality: Theory and applications* (R. Färe & D. Primont, Eds.). New York: Kluwer Academic Publishers.

Färe, R., & Zelenyuk, V. (2003). On aggregate Farrell efficiencies. *European Journal of Operational Research*, *146*(3), 615–620.

Fetter, R. B. (1991). Diagnosis related groups: Understanding hospital performance. *Interfaces*, *21*(1), 6–26.

Grosskopf, S. (1996). Statistical inference and nonparametric efficiency: A selective survey. *Journal of productivity analysis*, *7*(2-3), 161–176.

Grosskopf, S., Margaritis, D., & Valdmanis, V. (2001). Comparing teaching and non-teaching hospitals: A frontier approach (teaching vs. non-teaching hospitals). *Health Care Management Science*, *4*(2), 83–90.

Hao, S. S., & Pegels, C. C. (1994). Evaluating relative efficiencies of Veterans Affairs medical centers using Data Envelopment, ratio, and multiple regression analysis. *Journal of Medical Systems*, *18*(2), 55–67.

Härdle, W., & Simar, L. (2020). *Applied multivariate statistical analysis*. Springer.

Hollingsworth, B. (2003). Non-parametric and parametric applications measuring efficiency in health care. *Health Care Management Science*, *6*(4), 203–218.

Hollingsworth, B. (2008). The measurement of efficiency and productivity of health care delivery. *Health Economics*, *17*(10), 1107–1128.

Hu, H. H., Qi, Q., & Yang, C. H. (2012). Evaluation of China's regional hospital efficiency: DEA approach with undesirable output. *Journal of the Operational Research Society*, *63*(6), 715–725.

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

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

Kohl, S., Schoenfelder, J., Fügener, A., & Brunner, J. O. (2019). The use of data envelopment analysis (DEA) in healthcare with a focus on hospitals. *Health care management science*, *22*(2), 245–286.

Lee, C.-Y., & Cai, J.-Y. (2020). LASSO variable selection in data envelopment analysis with small datasets. *Omega*, *91*, 102019.

Li, Q. (1996). Nonparametric testing of closeness between two unknown distribution functions. *Econometric Reviews*, *15*(3), 261–274.

McCallion, G., McKillop, D. G., Glass, J. C., & Kerr, C. (1999). Rationalizing Northern Ireland hospital services towards larger providers: Best-practice efficiency studies and current policy. *Public Money & Management*, *19*(2), 27–32.

Meeusen, W., & van Den Broeck, J. (1977). Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, *18*(2), 435–444.

Mouchart, M., & Simar, L. (2002). *Efficiency analysis of air controllers: First insights* (Consulting report #0202). Institut de Statistique, Université Catholique de Louvain. Belgium.

Nguyen, B. H., & Zelenyuk, V. (2021a). Aggregate efficiency of industry and its groups: The case of Queensland public hospitals. *Empirical Economics*, forthcoming.

Nguyen, B. H., & Zelenyuk, V. (2021b). Robust efficiency analysis of public hospitals in Queensland, Australia. *Advances in contemporary statistics and econometrics* (forthcoming). Springer New York.

O'Neill, L., Rauner, M., Heidenberger, K., & Kraus, M. (2008). A cross-national comparison and taxonomy of DEA-based hospital efficiency studies. *Socio-Economic Planning Sciences*, *42*(3), 158–189.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, *27*(3), 832–837.

Sheather, S. J., & Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society. Series B (Methodological)*, *53*(3), 683–690.

Shephard, R. W. (1953). *Cost and production functions*. Princeton University Press.

Shephard, R. W. (1970). *Theory of cost and production functions*. Princeton University Press.

Sherman, H. D. (1984). Hospital efficiency measurement and evaluation: Empirical test of a new technique. *Medical Care*, *22*(10), 922–938.

Sickles, R., & Zelenyuk, V. (2019). *Measurement of productivity and efficiency*. Cambridge University Press.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Chapman; Hall.

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in Nonparametric Frontier Models. *Management Science*, *44*(1), 49–61.

Simar, L., & Wilson, P. W. (2000). A general methodology for bootstrapping in non-parametric frontier models. *Journal of applied statistics*, *27*(6), 779–802.

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

Simar, L., & Wilson, P. W. (2020). Hypothesis testing in nonparametric models of production using multiple sample splits. *Journal of Productivity Analysis*, *53*(3), 287–303.

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

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

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

Simar, L., & Zelenyuk, V. (2018). Central limit theorems for aggregate efficiency. *Operations Research*, *66*(1), 137–149.

Simar, L., & Zelenyuk, V. (2020). Improving finite sample approximation by central limit theorems for estimates from data envelopment analysis. *European Journal of Operational Research*, *284*(3), 1002–1015.

Staat, M. (2006). Efficiency of hospitals in Germany: A DEA-bootstrap approach. *Applied Economics*, *38*(19), 2255–2263.

Tiemann, O., & Schreyögg, J. (2009). Effects of ownership on hospital efficiency in Germany. *Academy of Management Proceedings*, *2009*(1), 1–6.

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

Zelenyuk, V. (2020). Aggregation of inputs and outputs prior to data envelopment analysis under big data. *European Journal of Operational Research*, *282*(1), 172–187.

Zhu, J. (1998). Data envelopment analysis vs. principal component analysis: An illustrative study of economic performance of Chinese cities. *European journal of operational research*, *111*(1), 50–61.

Zhu, J. (2020). DEA under big data: Data enabled analytics and network data envelopment analysis. *Annals of Operations Research*, in press.

Zuckerman, S., Hadley, J., & Lezzoni, L. (1994). Measuring hospital efficiency with frontier cost functions. *Journal of Health Economics*, *13*(3), 255–280.