Annals of Operations Research
https://doi.org/10.1007/s10479-023-05230-8

ORIGINAL RESEARCH

Predicting the performance of MSMEs: a hybrid DEA-machine learning approach

Sabri Boubaker[1,2,3] · **Tu D. Q. Le**[4,5] · **Thanh Ngo**[6,7] · **Riadh Manita**[8]

Accepted: 2 February 2023

Abstract

Micro, small and medium enterprises (MSMEs) dominate the business landscape and create more than half of employment worldwide. How we can apply big data analytical tools such as machine learning to examine the performance of MSMEs has become an important question to provide quicker results and recommend better and more reliable solutions that improve performance. This paper proposes a novel method for estimating a common set of weights (CSW) based on regression analysis for data envelopment analysis (DEA) as an important analytical and operational research technique, which (i) allows for measurement evaluations and ranking comparisons of the MSMEs, and (ii) helps overcome the time-consuming non-convexity issues of other CSW DEA methodologies. Our hybrid approach used several econometric and machine learning techniques (such as Tobit, least absolute shrinkage and selection operator, and Random Forest regression) to empirically explain and predict the performance of more than 5400 Vietnamese MSMEs (2010–2016), and showed that the machine learning techniques are more efficient and accurate than the econometric ones. Our study, therefore, sheds new light on the two-stage DEA literature, especially in terms of predicting performance in the era of big data to strengthen the role of analytics in business and management.

Keywords Machine learning (ML) · Common set of weights (CSW) · Data envelopment analysis (DEA) · Micro, small, and medium enterprise (MSME) · Efficiency

JEL Classification C61 · D24 · L60

✉ Thanh Ngo
T.ngo@massey.ac.nz

[1] EM Normandie Business School, Métis Lab, Paris, France

[2] International School, Vietnam National University, Hanoi, Vietnam

[3] Swansea University, Swansea, United Kingdom

[4] University of Economics & Law, Ho Chi Minh City, Vietnam

[5] Vietnam National University, Ho Chi Minh City, Vietnam

[6] School of Aviation, Massey University, Palmerston North, New Zealand

[7] VNU University of Economics & Business, Hanoi, Vietnam

[8] NEOMA Business School, Mont-Saint-Aignan, France

Published online: 14 February 2023

1 Introduction

Micro, Small and Medium Enterprises (MSMEs) play a key role in the global economy, accounting for about 90% of firms and creating more than 50% of employment worldwide (Ayyagari et al., 2003; IFC, 2012). In developing countries such as Vietnam, most MSMEs operate in the manufacturing sector (CIEM, 2016; GSO, 2016; Rand & Tarp, 2020), contributing to about 36% of the national value-added (OECD, 2021). It is therefore important to understand how efficiently MSMEs are operating and, especially, how to improve their performance. In the manufacturing sector, MSMEs are at the crossroads of technological advancement and operational excellence, where optimisation, Industry 4.0, and big data analysis are the buzzwords making the rounds (Schoenherr & Speier-Pero, 2015). A key research question arising from this situation is how to apply big data analytical tools such as machine learning (ML) to examine the performance of MSMEs, not only in terms of providing quicker results (regarding big data) but also in terms of recommending better and reliable solutions for improving their performance. Despite the growing body of literature on the application of analytics to solving operational problems (Kamble et al., 2020; Manimuthu et al., 2021; Wamba et al., 2017), research on MSMEs, especially in the manufacturing sector in developing countries, is still limited.

Data envelopment analysis (DEA) is a popular non-parametric tool for measuring efficiency and performance in various fields such as banking, healthcare, and aviation (Adler et al., 2002; Boubaker et al., 2018; Vidal-García et al., 2018; Yang, 2006). Zhu (2020) proposed that DEA should be viewed as a data-oriented analytical method for performance evaluations and benchmarking. The basic idea of DEA is that the individual decision-making unit (DMU) being examined can maximise its operational efficiency by using its own optimal weights regarding its inputs and outputs. The use of these so-called "dynamic weights" (Hammami et al., 2020) allows DEA to be price-free, and thus neither price information nor the functional form is needed. Consequently, DEA is more flexible with small samples, especially when the DMUs involved operate in a complex environment where it is difficult to define a production function (Ngo & Tsui, 2021). This has resulted in a much smaller number of DEA applications in the manufacturing sector (Tran & Ngo, 2014; Yang, 2006) where the data are large, especially given in big data era, normally involving thousands of DMUs or observations. Such studies often use the parametric approach of stochastic frontier analysis (Bačić et al., 2018; Hailu & Tanaka, 2015; Ngo et al., 2019a; Verschelde et al., 2016). One weakness of DEA, compared with stochastic frontier analysis, is that the different optimal weights allow the DMUs to be evaluated from different aspects, thus making it difficult to rank these DMUs on the same basis. The remedy to this situation is to estimate a common set of weights (CSW) that can be applied to all DMUs to provide a common basis for comparisons involving ranking; however, this approach incurs a high computational burden and sometimes faces the problem of convexity with non-linear objectives (Davtalab-Olyaie, 2019; Hammami et al., 2020; Wang et al., 2017, 2021).

DEA studies do not stop at the first stage of measuring efficiency; they also explain the role of environmental factors, including the corporate- and country-level governance of such efficiency in the second stage (Boubaker et al., 2019, 2020; Le et al., 2021). In other words, one may use the explanatory variables to explain or predict the efficiency scores of the DMUs involved. However, this second stage often applies the conventional econometric analytic models of Tobit or (bootstrap) truncated regression. According to Daraio et al. (2010, p. 1), "papers that estimate technical efficiency in the first stage and then regress these estimates on some environmental variables in a second-stage Tobit model continue to appear". In contrast, the more advanced estimators from ML such as Random Forest (RF) or neural network (NN) regressions have seldom been used. Since these ML estimators have better predictive power, they can overcome the problem of multicollinearity and are also tolerant to outliers and noise, and it is arguable that the application of such ML techniques can improve the explanatory or predictive results of two-stage DEA (Chen et al., 2021; Nandy & Singh, 2021; Thaker et al., 2021; Zhu et al., 2021).

Given the issues discussed above, the two specific research questions of this study were: "How can we efficiently measure the performance of Vietnamese manufacturing MSMEs using DEA but on the same basis?" and "How can we efficiently predict the performance of these MSMEs, given a set of corporate- and country-level variables?" For the former question, we need to have a novel CSW DEA model that can deal with big data, as this situation is problematic for both the dynamic weights and CSW DEA approaches. For the latter, we will need to compare several predictive methods, including both econometric and ML models. We expect to see the ML models perform better than the econometric ones.

This study, therefore, aimed to contribute to the literature in three aspects. First, we propose a novel method of estimating the CSW for measuring efficiency and comparing MSMEs by ranking via DEA. Since it is based on regression analysis (RA), this method helps overcome the time-consuming non-convexity issue of the previous CSW DEA methodologies. In this sense, it can be easily extended to other sectors where big data exist, thus widening the use of DEA in such studies. Second, by using data from more than 5400 Vietnamese manufacturing MSMEs operating during the 2010–2016 period, yielding a total of 37,557 observations, this study is among the first (DEA) studies focus on the performance of manufacturing MSMEs in developing countries to use big data. It is noted that the performance of manufacturing MSMEs has been examined in a few countries such as India (Kamble et al., 2020), Brazil (Borchardt et al., 2021), and Turkey (Sariyer et al., 2021), but a study combining big data and predictive analysis has not been conducted in the Vietnamese context. More importantly, the novel use of CSW means that it can provide widely acceptable recommendations for the MSMEs to help them improve their performance. Third, we used several econometric and ML techniques such as Tobit, the least absolute shrinkage and selection operator (LASSO), and RF regressions to compare their predictions regarding the performance of the examined MSMEs. Given the advantages of these ML techniques, our results are therefore more efficient and more accurate than the econometric ones. Consequently, our combined CSW DEA–ML approach can shed new light on the two-stage DEA literature, especially in terms of predicting performance and using big data and predictive analysis.

Empirically, our CSW–RA–DEA approach in the first stage showed that the Vietnamese MSMEs performed quite well during the 2010–2016 period, with the average efficiency scores consistently ranging from 0.803 to 0.824. Compared with the conventional DEA estimates, which ranged from 0.261 to 0.388, our results are more consistent with previous studies on manufacturing firms in Vietnam and other developing countries (Hailu & Tanaka, 2015; Le et al., 2018; Ngo et al., 2019a). Furthermore, the second-stage DEA on the determinants of such efficiency scores are also in line with the literature, in which the performance of Vietnamese MSMEs was negatively influenced by the firm's age, the ratio of female employees, and industrial zone status, but it was positively influenced by the firm's foreign ownership and participants, export activities, municipality status, the provincial business environment, and asset size. For big data and predictive analytical applications to predict the performance of these MSMEs, a hybrid approach of two popular econometric models from the DEA literature (namely Tobit and truncated regressions) and four ML algorithms (including LASSO, NN, support vector machine regression (SVR), and RF regression) were used in this study. Our findings suggest that the RF regression had the best in-sample predictive power (but this may have been caused by overfitting), the LASSO regression exhibited the best out-of-sample predictions, and the popular Tobit/truncated regressions were the worst performers for both in-sample and out-of-sample predictions. We argue that such econometric techniques are not suitable for predictive purposes, especially for big data.

We organised the rest of this article as follows. In the next section, we provide a brief discussion of DEA efficiency using the CSW, and the links between DEA and RA, as well as the increasing but limited uses of ML in DEA. Section 3 introduces the methodologies of conventional DEA and, more importantly, our novel CSW using RA in DEA (CSW-RA-DEA). Brief explanations of the ML techniques, including LASSO, SVR, and RF regressions, are also presented in this section. Section 4 then focuses on examining and predicting the performance of Vietnamese manufacturing MSMEs. Finally, Sect. 5 concludes the paper and suggests some directions for future research.

2 Literature review

2.1 DEA and the need for a CSW

It is acknowledged that DEA, which was developed by Charnes et al. (1978), is one of the most common methods used to evaluate efficiency in many fields (Contreras, 2020; Ngo & Tsui, 2021; Nguyen et al., 2019). Accordingly, the optimal weights can be used for the set of inputs and outputs, depending on the assumptions, which may be output-oriented, input-oriented, or even both (Hammami et al., 2020). This flexibility in the choice of weights may be both an advantage and a disadvantage of the method. When these weights are used, DEA becomes price-free, meaning that the relative efficiency of DMUs in the sample can be measured without the need for any functional form or price information (Contreras, 2020). However, different weights corresponding to different frontier surfaces could make it hard to compare and rank the DMUs, whether they are efficient or not (Jahanshahloo et al., 2008; Kao & Hung, 2005). Hence, variation in the optimal set of weights (the so-called “dynamic set of weights”) that is used to rank the DMUs may become inappropriate. This requires different ranking approaches.

The literature includes a number of ranking methods based on DEA, which can be divided into six groups (Adler et al., 2002) or 11 groups (Jahanshahloo et al., 2008). Most of them are based on the dynamic set of weights; therefore, comparing the DMUs among different frontier surfaces becomes an issue (Hammami et al., 2020). Kao and Hung (2005) emphasised that it is crucial to construct a CSW in DEA because a common frontier hyperplane will rank the DMUs according to the same aspect or criterion. In other words, the CSW will allow us to compare DMUs or the select the best DMU(s) in a fairer context (Contreras, 2020).

All CSW DEA involves two steps: (i) computing the DEA efficiency scores and the dynamic weights, then (ii) using optimisation, often as a programming problem with multiple objectives, to derive the CSW based on the dynamic weights (Davtalab-Olyaie, 2019; Wang & Chin, 2010; Wang et al., 2017, 2021), the efficiency distance (Kao & Hung, 2005; Wang et al., 2011), or the frontier distance (Hammami et al., 2020). This optimisation is time-consuming and sometimes faces the problem of convexity in the case of non-linear objectives. In line with the suggestion of Contreras (2020) that the CSW can be potentially determined by incorporating RA into DEA, this study proposed this use of RA to directly determine the CSW.

2.2 The integration of DEA, RA, and ML

The early work of Thanassoulis (1993) provided a comprehensive discussion comparing DEA and RA, and concluded that both methods can be used to complement each other where possible. Other authors also suggested that the corrected ordinary least squares frontier is analogous to DEA under the assumption of constant returns to scale (Greene, 2008). Ouenniche and Carrales (2018) further suggested that RA can provide DEA with feedback for variable selections in which the inputs (and outputs) are negatively (and positively) associated with the DEA efficiency scores.[1] In a similar vein, Tone and Tsutsui (2009) suggested that regression can be used to predict and adjust the data for multi-stage DEA. Furthermore, the CSW approaches of Kao and Hung (2005) and Wang et al. (2011), which aimed to minimise the efficiency distance (see Sect. 2.3)(Hammami et al., 2020), can be seen as a special case of RA (see Sect. 4)(Wang et al., 2011). Nonetheless, this reemphasises the importance of RA in DEA studies.

DEA studies therefore do not stop at the first stage of measuring efficiency. The role of environmental variables such as ownership, size, corporate governance, and other macro-economic factors can also be used in a second-stage regression to explain or predict the efficiency (Boubaker et al., 2019, 2020; Le et al., 2021). Since the DEA efficiency scores are bounded between 0 and 1, most of those studies used Tobit or truncated regressions (Daraio et al., 2010; Ho et al., 2021; Ngo et al., 2019b; Pilar et al., 2018). Given the big data era, there is an increasing but limited trend of using ML with DEA as a hybrid approach for analytical purposes (Khezrimotlagh et al., 2019). In particular, Zhu (2020) suggested that in big data situations, one needs to look at the possibility of combining DEA with other ML techniques, such as RF, support vector machines, and artificial neural networks. According to Tsai and Chen (2010) and Belhadi et al. (2021), among others, such hybrid combinations are superior to single models. For instance, Lee and Cai (2020) proposed using the least absolute shrinkage and selection operator (LASSO) for variable selection in DEA with small simulated datasets. Chen et al. (2021) extended this idea by using the elastic net (an extension of LASSO) in the more comprehensive setting of both small and big simulated data. Both studies showed that the hybrid approach performed better than the existing approaches. For second-stage regression, Wu et al. (2006) and Misiunas et al. (2016) demonstrated that an artificial NN can be trained by using data from the efficient DMUs; the results were used to adjust the dataset and selection of the variables to improve the predictive power of their DEA-NN model. Zhu et al. (2021) combined two ML techniques of NN and SVR into DEA to predict the efficiency scores even when new DMUs were added into the sample. Nandy and Singh (2021) and Thaker et al. (2021) relied on the use of RF regression to examine the impacts of the second-stage explanatory variables on the predicted efficiency scores, especially for an out-of-sample dataset. These studies also agreed on the superiority of the hybrid DEA-ML models. However, since all previous studies were based on the dynamic weights of DEA but not on the CSW, the impacts of these explanatory variables on the observed and predicted efficiency scores were not examined on the same basis. In this sense, this study filled this research gap by combining big data analytical tools (i.e., ML) and operational research approaches (i.e., DEA), and simultaneously accounted for the CSW when evaluating and predicting the performance of MSMEs.

[1] Specially, their findings showed that the DEA framework without feedback may lead to over- or underestimation of the efficiency scores (Contreras, 2020).

3 Methodologies

3.1 The research framework

This paper combines DEA analytics, econometric analytics, and ML analytics into a hybrid CSW-RA-DEA-ML predictive analytical method (see Fig. 1). Most ML studies forecast future outcomes on the basis of time series data; however, since our data spanned only 7 years (2010–2016), we did not have enough time-series datapoints for forecasting purposes. Instead, we focused on prediction, i.e., the use of pooled cross-sectional data, to answer our second research question about predicting the performance of Vietnamese MSMEs, given a set of corporate- and country-level information. To do so, our data were randomly split into two sub-samples, in which the training (in-sample) data consisted of 30,000 observations (about 80% of the total sample) and the predicting (out-of-sample) data consisted of 7557 observations (approximately 20% of the total sample).

Specifically, our study followed a three-stage analysis as described below.

First stage For the training data, we used CSW-RA-DEA (see Sect. 3.2) to estimate the efficiency of the 30,000 DMUs involved, using the firms' input and output data.

Second stage For the training data, we used different econometric (see Sect. 3.3) and ML (see Sect. 3.4) techniques to estimate the relationship between the CSW-RA-DEA efficiency (derived from stage 1) and the corporate- and country-level explanatory variables for the 30,000 DMUs involved, resulting in different predictive equations.

Third stage For the total sample, we used the predictive equations (derived in stage 2) to predict the DEA efficiency scores of all 37,557 DMUs involved, given the corporate- and country-level explanatory variables. Our estimates were then compared with the efficiency scores derived by traditional DEA in terms of the root mean squared error (RMSE). The technique or equation with the lowest RMSE exhibited the best predictive power.

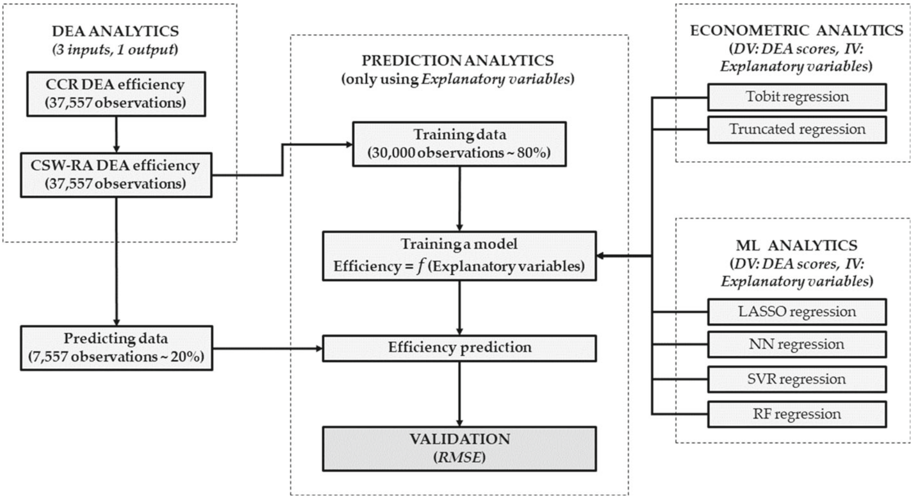

Fig. 1 The research framework

3.2 The DEA analytics

Consider a set of n DMUs, each using k inputs to produce m outputs. The goal of DEA is to estimate the optimal weights for the inputs or outputs for each DMU so that they can bring the DMU as close as possible to the frontier envelope of the DMUs (i.e., to maximise the DMU's efficiency). The mathematical expression of (constant returns to scale) DEA,[2] as introduced by Charnes et al. (1978), is:

$$EF_{j_0} = max_{u,v} \frac{\sum_{r=1}^{m} u_r y_{rj_0}}{\sum_{i=1}^{k} v_i x_{ij_0}} \tag{1}$$

Subject to

$$\frac{\sum_{r}^{m} u_r y_{rj}}{\sum_{i}^{k} v_i x_{ij}} \leq 1, \forall j, \quad j = 1, 2, \ldots, n$$
$$u_r, v_i \geq \varepsilon, \quad \forall i, r$$

where θ_{j_0} is the efficiency score of DMU j_0 ($j = 1,2,\ldots,n$) to be maximised, given the output weight u_r of output y_r ($r = 1,2,\ldots,m$) and the input weight v_i of input x_i ($i = 1,2,\ldots,k$); ε is a non-Archimedean value designed to ensure positive weights. It is noted that Eq. (1) needs to be run n times for each of the DMUs in the sample, in which the optimal weights (v_i, u_r) can vary among the DMUs; hence, the so-called argument of the "dynamic set of weights" in DEA.

The CSW DEA seeks a common set of weights that can be applied to all DMUs in the sample instead of using different weights for each DMU. This makes sense for managers and decision-makers because it helps in benchmarking and ranking the DMUs in the same terms so that any recommendations or policies can be widely accepted and feasibly applied by these DMUs. Unlike previous CSW DEA studies, however, this study proposed the novel approach of CSW–RA–DEA to determine the CSW. The algorithm of our CSW-RA-DEA approach is described below.

Step 1 Compute the DEA efficiency scores for all DMUs in the sample as normal via Eq. (1). Note that all CSW DEA studies have applied this step.

Step 2 Regress those efficiency scores on the inputs x_i and outputs y_r of the DMUs. According to Ouenniche and Carrales (2018), among others, all inputs need to be negatively associated with the efficiency scores, whereas the relationship between the efficiency scores and the outputs should be positive. This regression has the form:

$$EF_j = \alpha_0 + \beta_i x_{ij} + \gamma_r y_{rj} + \varepsilon \tag{2}$$

where β_i is expected to be negatively significant and γ_r is expected to be positively significant.

[2] Although there have been various improvements and extensions of DEA (e.g. Andersen and Petersen, 1993; Banker, 1984; Davtalab-Olyaie, 2019; Ngo and Tsui, 2021) – and the readers are encouraged to find more information therein – this study only focused on the traditional DEA model of Eq. (1) because of its simplicity.

Step 3 The CSW will be $(-\beta_i, \gamma_r)$, with a negative sign on β_i to convert it to a positive value; accordingly, the CSW-RA-DEA efficiency scores can be estimated as[3]:

$$\text{(CSW - RA - DEA)}\ EF_j = \frac{\sum_{r=1}^{m} \gamma_r y_{rj}}{\sum_{i=1}^{k} \beta_i x_{ij}} \tag{3}$$

We also used two popular numerical examples in the CSW literature (Davtalab-Olyaie, 2019; Sexton et al., 1986; Wang & Chin, 2010; Wang et al., 2021) to compare the efficiency scores and the ranks derived by different CSW approaches, including our CSW–RA–DEA approach. Our results show that the CSW–RA–DEA approach provided consistent and even better results than the others (see the Appendix) and that, therefore, it was appropriate to use in our analysis.

3.3 The econometric analytics

Obviously, one can train a model to estimate the impacts of the explanatory variables on the dependent variable, such as the CSW–RA–DEA efficiency scores derived from the DEA approach, following traditional econometric approaches. Since the efficiency scores are bounded between 0 and 1, it can be argued that Tobit or truncated regression is more appropriate for this second-stage DEA (Boubaker et al., 2019; Ho et al., 2021; Ngo et al., 2019b). For example, a simple search on Google Scholar on 20 November 2021 with the keywords “DEA”, “efficiency”, “Tobit”, and “two stage” returned 6540 results; a similar search using the keywords “DEA”, “efficiency”, “truncated”, and “two stage” resulted in 4680 results. Both models have the form:

$$EF = \alpha + \beta Z + \epsilon \tag{4}$$

in which

$$EF = \left\{ \begin{array}{lll} 0 & \text{if} & EF < 0 \\ EF & \text{if} & 0 \leq EF \leq 1 \\ 1 & \text{if} & EF > 1 \end{array} \right\} \tag{5}$$

where EF is the CSW–RA–DEA efficiency scores, Z is the vector of the explanatory variables, β is the vector of the coefficients to be estimated, α is the intercept, and ϵ is the random error.

3.4 The ML analytics

The current big data era is witnessing a growing body of literature on the use of ML for prediction purposes (Manimuthu et al., 2021; Schoenherr & Speier-Pero, 2015; Wamba et al., 2017; Zhu et al., 2021;), particularly in DEA studies (Nandy & Singh, 2021; Thaker et al., 2021). We therefore follow the literature in using LASSO regression (Chen et al., 2021), NN regression (Wu et al., 2006; Zhu et al., 2021), SVR (Zhu et al., 2021), and RF regression (Nandy & Singh, 2021; Thaker et al., 2021) as the ML analytical techniques for training our prediction model. This section briefly introduces these ML algorithms; the readers are encouraged to find more technical information in the relevant literature and the references therein.

[3] Equation (3) has no constraint to limit the CSW-RA’s efficiency score to less than 1; we can treat these scores as being similar to the case of super-efficiency (Andersen and Petersen, 1993).

LASSO identifies the important explanatory variables of the dependent efficiency scores by minimizing the following L1 penalization on the total sum of coefficients (Lee & Cai, 2020):

$$\min_{\alpha,\beta} \frac{1}{2} \sum_{j=1}^{n} (EF - \alpha - \beta Z)^2 + \lambda \sum |\beta| \tag{6}$$

where λ is a penalty (or tuning parameter) chosen by the extended Bayesian information criterion. Note that when $\lambda = 0$, the LASSO model in (6) collapses into the traditional regression model in (4). It is noted that, as is the case in other ML algorithms, including the other ones presented in this section, the estimation of the vector β in LASSO does not focus on its significance; it focuses on the contributions of each explanatory variable to the construction or prediction of the efficiency scores instead.

The NN model provides a different method that uses hidden layers to extract the important features or inputs of a given output (Wu et al., 2006; Zhu et al., 2021). In our case, it was appropriate to extract the important explanatory variables (inputs) influencing the CSW–RA–DEA efficiency scores (output) using NN. Specifically, the NN algorithm started by estimating the weights (or importance) of the inputs, then the relevant output was mapped via an activation function $f(\bullet)$. This output was compared with the desired output, and the error was calculated accordingly. The error was then back-propagated to the NN to help adjust the weights, with the aim of reducing the error in each iteration. In our study, the activation function $f(\bullet)$ had the form:

$$EF = f\left(\sum wZ\right) - \theta \tag{7}$$

where $\sum wZ$ is the weighted sum of the explanatory variables (or inputs) Z and θ is the intercept.

RF is another ML algorithm which is based on a decision tree (Breiman et al., 1984) and the bootstrapping (or “bagging”) technique (Breiman, 1996). It randomly bootstraps the training dataset many times; at each iteration, the data are recursively partitioned by one input at a time (also called a “node”) to create a decision tree. By combining all these random “trees”, RF generates a “forest” where the dependent output can be predicted as the average of the predictions of all trees (Nandy & Singh, 2021; Thaker et al., 2021). According to Thaker et al. (2021), the RF’s predictor is computed as:

$$EF = \frac{1}{B} \sum_{b=1}^{B} Q(Z, \Theta_b) \tag{8}$$

where B is the number of randomised trees in the forest (i.e., the number of bootstrap iterations) and Q represents the predicted output of each tree, given the input Z and the independent and identically distributed random vector Θ_b that represents the relationship between the inputs and the output in the tree of the b-th iteration.

The algorithm of SVR is slightly different from the ones described above. Instead of examining the relationship between the inputs and the output, SVR constructs a hyperplane to separate the data, given the multiple-dimensional space of the output and inputs. In other words, the aim of SVR is to find the optimal surface that minimises the error of all training datapoints on the hyperplane (Smola & Schölkopf, 2004; Zhu et al., 2021). The linear form of SVR is:

$$EF = wZ + b \tag{9}$$

where w represents the support vectors of the hyperplane and b is the intercept, which are the optimal solutions of:

$$\min_{w,b,\xi_j,\xi_j^*} \frac{1}{2}w^2 + C\sum_{j=1}^{n}\left(\xi_j + \xi_j^*\right) \tag{10}$$

Subject to

$$EF_j - (wZ + b) \le \epsilon + \xi_j^*,\ \ j = 1, 2, \ldots n$$

$$(wZ + b) - EF_j \le \epsilon + \xi_j,\ \ j = 1, 2, \ldots n$$

$$\xi_j, \xi_j^* \ \ge\ 0, \forall j$$

4 Analytics using CSW–RA–DEA: the performance of Vietnamese manufacturing MSMEs

This section provides an analytical application of CSW–RA–DEA to a rich dataset of more than 37,000 observations on MSMEs in the Vietnamese manufacturing industry during 2010–2016. Given the rising use of big data, as in our case, defining the CSW via previous approaches by linear or non-linear optimisation of the secondary goal (Davtalab-Olyaie, 2019; Wang & Chin, 2010; Wang et al., 2021) is time-consuming but this was justified for our proposed CSW-RA-DEA approach.

4.1 Data and variable selection

Vietnam is an emerging economy that has witnessed impressive economic development over the last few decades. The driving force behind its economic growth is household businesses or MSMEs (CIEM, 2016; OECD, 2021; Rand & Tarp, 2020). For instance, Rand and Tarp (2020) emphasised that in Vietnam, private SMEs accounted for about 95% of all enterprises, employed about half of the workforce, and produced approximately 40% of the national GDP. Given the key role that the MSMEs play nationally and globally (Ayyagari et al., 2003; IFC, 2012; OECD, 2021), it is therefore important to examine the performance and efficiency of Vietnamese MSMEs, especially for making important recommendations to managers and policymakers to help improve the performance of this sector. Importantly, with data on Vietnamese MSMEs comprising more than 37,000 firm-year observations, this sample was suitable for a hybrid study combining DEA and ML techniques.

In line with the literature on evaluating the efficiency of manufacturing firms, such as Verschelde et al. (2016), Ngo et al. (2019a), and Sahoo et al. (2021), we examined the Vietnamese MSMEs in terms of three important inputs, namely labour (proxied by the number of employees, x_1), capital (proxied by the value of total assets, x_2), and materials (proxied by the amount of materials, x_3), to produce a single output (total revenue, y). This information was extracted from the annual surveys of Vietnamese enterprises conducted by the national General Statistics Office (GSO, 2016); such data are popular in many studies (Dao et al., 2021; Le et al., 2018; Rand & Tarp, 2020). Since we focused our study on MSMEs only, we followed the IFC (2012) and filtered out firms with more than 250 employees, resulting in 37,557 firm-year observations for the Vietnamese MSMEs operating during the 2010–2016 period.

Accordingly, our data covered 2011 observations for micro (one to nine employees), 12,494 observations for small (10–49 employees), and 23,052 observations for medium (50–249 employees) enterprises. Note that most previous studies applied a two-stage analysis, where the (dynamic) DEA efficiency scores were estimated (in the first stage) then regressed on a set of explanatory variables (the second stage).[4] More importantly, if these explanatory variables were found to significantly influence the CSW–RA–DEA efficiency, we could use them to predict the performance of the Vietnamese MSMEs. We therefore used several prediction methods, including recent ML techniques such as LASSO and RF regressions, in our second-stage analysis. The basic information of our data and variables are presented in Table 1.

4.2 First-stage analytics: the CSW–RA–DEA efficiency of Vietnamese MSMEs

We report our CSW-RA-DEA efficiency scores for our sample of 37,557 MSME observations in Table 2, in which the estimated CSW–RA–DEA scores have higher means than the (dynamic) DEA scores. On the one hand, we can see that the average CSW-RA-DEA efficiency scores, which consistently ranged from 0.803 to 0.824 during the 2010–2016 period, in agreement with previous studies on manufacturing firms in Vietnam and other developing countries (Hailu & Tanaka, 2015; Le et al., 2018; Ngo et al., 2019a), compared with traditional DEA scores. On the other hand, we argue that the use of RA (in step 2) allowed us to estimate the weighted inputs and outputs (in step 3 of the CSW-RA-DEA algorithm) as having greater variations; therefore, the CSW–RA–DEA efficiency scores can have a wider range. However, this is similar to the case of super-efficiency (see Sect. 3.3 above) or other econometric-based DEA results (Wu et al., 2006). More importantly, the results of both Spearman's and Kendall's ranking correlations in Table 3 confirmed that our CSW-RA-DEA estimations are consistent with the results of traditional DEA and thus are reliable. In this sense, it was justified to proceed with the second-stage regression.

4.3 Second-stage analytics: predicting the Vietnamese MSMEs' performance

The predictions of our econometric and ML analytics are presented in Table 4. Three important findings and their relevant managerial implications can be summarised as follows.

Firstly, from the managerial perspective, Table 4 suggests (and confirms) that the performance of Vietnamese MSMEs was (i) negatively influenced by the firm's age, the ratio of female employees, and the industrial zone status; and (ii) positively influenced by the firm foreign ownership, export activities, the municipality status, the provincial business environment, and asset size. These findings are consistent with the literature. For instance, it can be argued that young firms are more likely to be involved with radical innovations (Acemoglu & Cao, 2015); unlike in other sectors, where the use of technology and innovations may be an obstacle (Pellegrino, 2018). For MSMEs, such innovations do not require many resources and are feasible. Because most Vietnamese MSMEs operate in the garment, textile, and footwear sector (Dao et al., 2021; Pham et al., 2010), where the productivity of female employees is still low, it is reasonable to see that firms with a higher female employee ratio tend to have lower efficiency. In contrast, the participation of foreign investors allows the firms to possess more advanced technologies and management knowledge and hence, improve their

[4] These explanatory variables are deemed to affect firm efficiency. For more details, see Anh and Gan (2020), Dao et al. (2021), and Sahoo et al. (2021).

Table 1 Descriptions of the variables

Variable	Mean	Definition	Previously used in
First-stage: CSW-RA-DEA			
x_1	85.32	Number of employees	Tran and Ngo (2014), Hailu and Tanaka (2015), Verschelde et al. (2016), Pilar et al. (2018), Ngo et al. (2019a) and Anh and Gan (2020)
x_2	33,343.42	Value of total assets (million VND)	
x_3	588.47	Value of materials used (million VND)	
y	74,098.85	Total revenue (million VND)	
Second-stage: Econometric and ML analyses			
AGE	11.53	Number of years in operation	Vu et al. (2016), Bačić et al. (2018), Le et al. (2018), Pilar et al. (2018), Nguyen et al. (2019), Anh and Gan (2020) and Sahoo et al. (2021)
SOE	0.06	Dummy variable that equals 1 if the firm is a central or local state-owned company	
FOE	0.57	Dummy variable that equals 1 if the firm is 100% foreign-owned or is a joint venture with foreign capital	
FERATIO	0.43	Ratio of female employees to total employees	
EX	0.50	Dummy variable that equals 1 if the firm is involved in exporting activities	
IZONE	0.64	Dummy variable that equals 1 if the firm is located inside an industrial zone	
MUNI	0.25	Dummy variable that equals 1 if the firm's headquarter is located in one of the five municipalities (Hanoi, Hochiminh City, Haiphong, Danang, and Can Tho)	
PCI	59.49	Provincial competitiveness index	
SIZE	10.67	The logarithmic value of total assets	
TECH	0.19	Dummy variable that equals 1 if the firm operates in a high-tech industry involving chemicals, pharmaceuticals, computers, machinery, vehicles, and/or equipment	
T	4.00	Time variable: 1 = 2010, 2 = 2011, and so on	
MICRO	0.05	Dummy variable that equals 1 if the firm has fewer than 10 employees	
SMALL	0.33	Dummy variable that equals 1 if the firm has 10–49 employees	
MEDIUM	0.61	Dummy variable that equals 1 if the firm has more than 50 employees	

The total number of observations is 37,557

Table 2 Average efficiency scores of DEA and CSW-RA-DEA for 37,557 MSME observations (2010–2016)

	Obs	Mean	SD	Minimum	Maximum
2010					
DEA	5,366	0.267	0.119	0.045	1.000
CSW–RA–DEA	5,366	0.817	0.369	0.033	5.487
2011					
DEA	5,370	0.334	0.162	0.024	1.000
CSW–RA–DEA	5,370	0.824	0.362	0.024	4.819
2012					
DEA	5,355	0.388	0.152	0.019	1.000
CSW–RA–DEA	5,355	0.803	0.356	0.015	3.136
2013					
DEA	5,415	0.362	0.162	0.031	1.000
CSW–RA–DEA	5,415	0.815	0.359	0.014	3.653
2014					
DEA	5,305	0.333	0.147	0.030	1.000
CSW–RA–DEA	5,305	0.814	0.359	0.031	3.419
2015					
DEA	5,366	0.261	0.141	0.001	1.000
CSW–RA–DEA	5,366	0.817	0.371	0.001	5.649
2016					
DEA	5,380	0.349	0.154	0.032	1.000
CSW–RA–DEA	5,380	0.808	0.359	0.035	2.901
Whole period (2010–2016)					
DEA	37,557	0.328	0.155	0.001	1.000
CSW–RA–DEA	37,557	0.814	0.362	0.001	5.649

DEA, efficiency scores estimated under the assumption of constant returns to scale; CSW-RA-DEA, efficiency scores estimated by our approach; Obs, number of observations; SD, standard deviation

Table 3 Ranking correlations between DEA and CSW-RA-DEA scores

	2010	2011	2012	2013	2014	2015	2016	2010–2016
Spearman	0.565	0.696	0.781	0.709	0.863	0.589	0.912	0.673
Kendall	0.420	0.538	0.609	0.548	0.689	0.440	0.797	0.503

All figures are significant at the 1% level

performance (Huang & Yang, 2016; Ngo et al., 2019a). Similarly, MSMEs involved in export activities can benefit from a learning-by-exporting effect because they are more exposed to foreign technology and competition (Amiti & Konings, 2007; Baldwin & Gu, 2004; Pilar et al., 2018). The MSMEs also benefit from operating in large municipalities because of the effects of firm selection and agglomeration economies (Combes et al., 2012; Le et al., 2018; Vu et al., 2016), having a good provincial business environment (Ngo et al., 2019a;

Table 4 Predictive performance of different methods

	TOBIT		TRUNCATED		LASSO	NN	RF
	Coefficient		Coefficient		Coefficient	Coefficient	Variable importance
AGE	– 0.0007	***	– 0.0003		– 0.0007	0.0010	0.4165
SOE	– 0.0201	**	– 0.0183	*	– 0.0112	– 0.0005	0.3109
FOE	0.0592	***	0.0642	***	0.0618	0.0221	0.6331
FERATIO	– 0.0356	***	– 0.0282	***	– 0.0316	– 0.0070	0.5067
EX	0.0206	***	0.0181	***	0.0200	0.0015	0.4163
IZONE	– 0.0207	***	– 0.0159	***	– 0.0169	– 0.0137	0.3313
MUNI	0.0625	***	0.0915	***	0.0557	0.0116	0.5795
PCI	0.0080	***	0.0056	***	0.0077	0.1113	0.5010
SIZE	0.0118	***	0.0064	***	0.0115	0.1056	0.5231
TECH	0.0146	***	0.0144	**	0.0143	0.0024	0.3327
T	– 0.0002		– 0.0004			0.0021	0.3472
MICRO	– 0.1575	***	– 0.1545	***	– 0.1374	– 0.0076	1.0000
SMALL	– 0.0235	***	– 0.0236	***	– 0.0193	– 0.0062	0.3346
CONSTANT	0.2164	***	0.3057	***	0.2283		
Explanatory variables	13		13		12	13	13
In-sample RMSE	0.51161		0.36900		0.35610	0.35410	**0.13713**
Out-of-sample RMSE	0.51400		0.35973		**0.34627**	0.35129	0.34858

AGE, number of years of operation; SOE, dummy variable for state ownership; FOE, dummy variable for foreign ownership; FERATIO, ratio of female employees to total employees; EX, dummy variable for exporting activities; IZONE, dummy variable for industrial zone status; MUNI, dummy variable for municipality status; PCI, provincial competitiveness index; SIZE, logarithmic value of total assets; TECH, dummy variable for high-tech industries (chemicals, pharmaceuticals, computers, machinery, vehicles, and equipment); T, a continuous time variable which equals 1 for the year 2010 and so on; MICRO, dummy variable for micro enterprises; SMALL, dummy variable for small enterprises; TOBIT, Tobit regression; TRUNCATED, truncated regression; LASSO, LASSO regression; NN, neural network regression; RF, random forest regression; RMSE, root mean square error. The lower the RSME, the more accurate the prediction model

***, **, and * denote the 1%, 5%, and 10% level of significance, respectively

Dao et al., 2021; VCCI & USAID, 2022), and economies of scale (Bačić et al., 2018; Ngo et al., 2019a), to further improve their efficiency. These findings were robust across different models, including Tobit, truncated, and LASSO regressions. The RF regression did not provide coefficient estimates but confirmed the contributions of these explanatory variables: for instance, it identified MICRO as the most important factor (variable importance = 1.000), as it had the greatest magnitude for its coefficients: – 0.1575, – 0.1545, and – 0.1374 in the Tobit, truncated, and LASSO regressions, respectively. The NN model supported most of the signs of the coefficients but not their significance; however, this model performed slightly worse than the other ML models (the in-sample and out-of-sample RMSEs were slightly high at 0.35410 and 0.35129, respectively).

Secondly, we observed the disadvantages in terms of the competitiveness and performance of micro and small enterprises, compared with medium ones (Kamble et al., 2020), with the coefficients of both MICRO and SMALL being negatively and statistically significant. We also found that high-tech MSMEs outperformed their counterparts, in line with the evidence provided by Anh and Gan (2020). We therefore suggest that Vietnamese manufacturing MSMEs should be encouraged to expand their scale (both in terms of assets and employment) and become more involved in export activities. Meanwhile, central and provincial Vietnamese governments should improve their (business) governance and to allow more activities and the involvement of foreign investors in the Vietnamese manufacturing sector. As discussed earlier, given that our estimates are based on the CSW–RA–DEA scores, we believe that these managerial suggestions and recommendations can be widely applied to all MSMEs in our sample.

Thirdly, from the methodological perspective, the last three rows of Table 4 suggest that the advanced ML analysis generally made better predictions for both in-sample and out-of-sample data (i.e., lower RMSEs) compared with the traditional econometric analytical methods (i.e., Tobit and truncated regressions). Among the ML techniques, LASSO regression was the best model for out-of-sample prediction. The RF model seemed to be overfitted for the in-sample data (with an exceptional low in-sample RMSE of only 0.1371), but its predictive ability for out-of-sample data was not remarkable (its out-of-sample RMSE was 0.34858). Although this is not reported in Table 4, a similar situation was found for SVR, for which the in-sample and out-of-sample RMSEs were 0.31632 and 0.36842, respectively. For the econometric models, truncated regression outperformed the Tobit model, supporting the argument of Daraio et al. (2010) that DEA efficiency scores are truncated rather than censored. Nevertheless, both models yielded high RMSE values. Therefore, although the two are popular for two-stage DEA studies, we suggest that they are not suitable for predictive purposes.[5] We argue that for big data samples such as in our case, censoring or truncating efficient observations (with DEA scores greater than or equal to 1) from the prediction model may result in missing information, and their predictive power is accordingly weaker. We therefore support the ML literature (e.g., Belhadi et al., 2021; Tsai & Chen, 2010; Zhu et al., 2021) in confirming that the ML approach is superior to the econometric approach, and that our hybrid DEA-ML model is more efficient and more accurate than the traditional ones. Consequently, we conclude that LASSO regression is the best model of the ML approaches for predicting the efficiency of Vietnamese MSMEs.

Nevertheless, we have shown that the CSW–RA–DEA yields consistent and better results than other CSW ranking methods such as cross-efficiency, super-efficiency, normalised common weights, and so on. Because our novel model is based on RA, it overcomes the time-consuming non-convexity issue of previous CSW DEA methodologies, especially for large samples. As such, the CSW–RA–DEA model could be extended to other fields where big data exist, thus widening the use of DEA in such fields.

5 Conclusions

This study proposed a novel method of estimating the common set of weights for evaluating performance and rankings via DEA based on regression analysis (CSW–RA–DEA). It then applied CSW–RA and several other prediction methods (two econometric models and four ML algorithms) to explain and predict the performance of more than 5400 Vietnamese manufacturing MSMEs operating during 2010–2016. In this sense, our study contributes to

[5] To check the robustness, we experimented with (i) using the original DEA efficiency scores as the dependent variable instead of the CSW-RA-DEA scores, and (ii) using the normalized values of the CSW–RA–DEA scores to avoid too many scores being greater than 1. Both cases resulted in the conclusion that ML is more suitable than Tobit or truncated regressions in terms of predictive ability.

the literature in terms of methodological (the CSW–RA–DEA method itself as well as the hybrid DEA-ML approach), empirical (the use of CSW DEA for Vietnamese MSMEs), and managerial (recommendations for improving MSMEs' performance) perspectives.

It should be noted that although we have examined four popular ML techniques in our hybrid DEA–ML model (i.e., LASSO, NN, SVR, and RF regressions), there are other ML algorithms and DEA models that could be investigated in future research. Regarding the DEA approach, one could apply the variable returns to scale assumption (Banker, 1984), the cost/profit measures (Ngo et al., 2019b; Pilar et al., 2018), the fuzzy approach (Boubaker et al., 2020; Nandy & Singh, 2021), or the Euclidean distance (Hammami et al., 2020) to measure the DEA efficiency in different settings. Regarding ML analytics, the ensemble approach (Belhadi et al., 2021) and other hybrid ML combinations (e.g., combining LASSO and NN) could also be used. Finally, the extension of this CSW–RA–DEA to other industries with big data, such as banking and finance (Tsai & Chen, 2010), healthcare (Misiunas et al., 2016), agriculture (Nandy & Singh, 2021), or energy (Khezrimotlagh et al., 2019), could help increase our understanding of the role of DEA in such fields.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions.

Declarations

Conflict of interest The authors declare no conflict of interest.

Appendix The consistency of CSW-RA efficiency scores

This appendix provides a comparison between the results of CSW–RA and the results from other CSW approaches by applying two numerical examples popularly used in the literature.

Example 1: We started with the simple numerical example of six nursing homes with two inputs (staff hours per day, x_1; suppliers per day, x_2) to produce two outputs (total Medicare and Medicaid-reimbursed patient days, y_1; total privately paid patient days, y_2). This example has been examined in Sexton et al. (1986) and Wang et al. (2011), and the input/output data are presented in Table 5. Table 5 also compares our CSW-RA-DEA efficiency scores with those derived by traditional DEA (Charnes et al., 1978), the super-efficiency approach (Andersen & Petersen, 1993), the cross-efficiency approach (Sexton et al., 1986), the compromise solution approach (Kao & Hung, 2005), and the RAM approach (Wang et al., 2011). As observed in Table 5, similar to the super-efficiency approach, the RAM and CSW–RA–DEA approaches can provide a full ranking comparison for all DMUs but the cross-efficiency and compromise solution approaches cannot. More importantly, CSW–RA–DEA ranked the examined DMUs better than RAM, regarding the super-efficiency ranking, suggesting that CSW–RA–DEA is

Table 5 Example of six nursing homes

DMU	x_1	x_2	y_1	y_2	DEA		Sup-EF		Cross–EF		CSA–EF		RAM–EF		CSW–RA–DEA	
A	150	0.2	14,000	3500	1.000	(1)	2.000	(1)	0.939	(5)	1.000	(1)	1.142	(1)	1.424	(1)
B	400	0.7	14,000	21,000	1.000	(1)	1.400	(3)	0.946	(1)	0.920	(5)	0.973	(5)	1.126	(2)
C	320	1.2	42,000	10,500	1.000	(1)	1.406	(2)	0.939	(5)	0.924	(4)	0.985	(4)	1.101	(3)
D	520	2	28,000	42,000	1.000	(1)	1.131	(4)	0.946	(1)	1.000	(1)	1.003	(2)	1.068	(4)
E	350	1.2	19,000	25,000	0.977	(5)	0.977	(5)	0.946	(1)	0.972	(3)	0.986	(3)	1.065	(5)
F	320	0.7	14,000	15,000	0.867	(6)	0.867	(6)	0.946	(1)	0.831	(6)	0.875	(6)	0.996	(6)

x_1, staff hours per day; x_2, suppliers per day; y_1, total Medicare and Medicaid-reimbursed patient days; y_2, total privately paid patient days; DEA, efficiency scores estimated under the assumption of constant returns to scale; Sup-EF, super-efficiency; Cross-EF, cross-efficiency scores reported as EROW in Table 8 of Sexton et al. (1986); CSA-EF, efficiency scores estimated by the compromise solution approach of CSW as reported in Table 5 of Wang et al. (2011); RAM-EF, efficiency scores estimated by the regression analysis method of CSW as reported in Table 5 of Wang et al. (2011); CSW–RA–DEA, efficiency scores estimated by our approach. The ranks of the DMUs are presented inside the brackets

Table 6 Spearman's ranking correlations among the estimated CSW efficiency scores for six nursing homes

	DEA	Sup-EF	Cross-EF	CSA–EF	RAM–EF	CSW–RA–DEA
DEA	**1.000**					
	(0.000)					
Sup-EF	0.845	**1.000**				
	(0.034)	(0.000)				
Cross-EF	– 0.490	– 0.828	**1.000**			
	(0.324)	(0.042)	(0.000)			
CSA–EF	0.515	0.464	– 0.315	**1.000**		
	(0.296)	(0.354)	(0.543)	(0.000)		
RAM–EF	0.507	0.543	– 0.414	0.986	**1.000**	
	(0.305)	(0.266)	(0.414)	(0.000)	(0.000)	
CSW–RA	0.845	0.943	– 0.621	0.406	0.486	**1.000**
	(0.034)	(0.005)	(0.188)	(0.425)	(0.329)	(0.000)

p-values are presented inside the brackets

the best ranking approach among these methods. The strong relationship among the results of CSW–RA–DEA, DEA, and Sup-EF in Table 6 further supports this argument.

Example 2: This example involves 14 international passenger airlines with three inputs (aircraft capacity, x_1; operating costs, x_2; non-flight assets, x_3) and two outputs (passenger-kilometres,y_1; non-passenger operating revenue, y_2). These data have been used in Wang and Chin (2010), Wang et al. (2017), and Davtalab-Olyaie (2019), who applied the cross-efficiency approach, and by Wang et al. (2021), who applied a normalized weights approach. Again, in Tables 7 and 8, we can observe a highly consistent relationship between the efficiency scores and rankings across different CSW approaches, confirming the ranking ability and the usefulness of the CSW–RA–DEA approach.

Table 7 Example of 14 international airlines

DMU	x1	x2	x3	y1	y2	DEA		Sup-EF		Cross-EF		NCross-EF		Novel-CSW		NCSW		CSW–RA–DEA	
1	5,723	3,239	2,003	26,677	697	0.868	(12)	0.868	(12)	0.752	(12)	0.705	(11)	0.797	(10)	0.819	(12)	0.916	(12)
2	5,895	4,225	4,557	3,081	539	0.338	(14)	0.338	(14)	0.190	(14)	0.191	(14)	0.184	(14)	0.201	(14)	0.206	(14)
3	24,099	9,560	6,267	124,055	1,266	0.947	(11)	0.947	(11)	0.768	(9)	0.715	(10)	0.820	(9)	0.865	(9)	0.994	(7)
4	13,565	7,499	3,213	64,734	1,563	0.958	(9)	0.958	(9)	0.820	(6)	0.773	(7)	0.863	(6)	0.891	(6)	0.985	(8)
5	5,183	1,880	783	23,604	513	1.000	(1)	1.221	(4)	0.896	(3)	0.876	(2)	0.935	(4)	0.967	(4)	1.128	(2)
6	19,080	8,032	3,272	95,011	572	0.977	(8)	0.977	(8)	0.755	(11)	0.702	(12)	0.787	(12)	0.846	(10)	0.941	(9)
7	4,603	3,457	2,360	22,112	969	1.000	(1)	1.021	(6)	0.815	(7)	0.771	(8)	0.840	(7)	0.868	(8)	0.931	(11)
8	12,097	6,779	6,474	52,363	2,001	0.859	(13)	0.859	(13)	0.723	(13)	0.691	(13)	0.757	(13)	0.780	(13)	0.881	(13)
9	6,587	3,341	3,581	26,504	1,297	0.948	(10)	0.948	(10)	0.759	(10)	0.738	(9)	0.793	(11)	0.821	(11)	0.933	(10)
10	5,654	1,878	1,916	19,277	972	1.000	(1)	1.268	(3)	0.786	(8)	0.781	(6)	0.836	(8)	0.871	(7)	1.022	(6)
11	12,559	8,098	3,310	41,925	3,398	1.000	(1)	1.896	(1)	0.919	(1)	0.904	(1)	0.958	(2)	1.000	(1)	1.052	(5)
12	5,728	2,481	2,254	27,754	982	1.000	(1)	1.007	(7)	0.887	(4)	0.854	(4)	0.944	(3)	0.972	(3)	1.127	(3)
13	4,715	1,792	2,485	31,332	543	1.000	(1)	1.381	(2)	0.919	(2)	0.872	(3)	0.985	(1)	1.000	(1)	1.216	(1)
14	22,793	9,874	4,145	122,528	1,404	1.000	(1)	1.092	(5)	0.866	(5)	0.814	(5)	0.908	(5)	0.935	(5)	1.077	(4)

x_1, aircraft capacity; x_2, operating costs; x_3, non-flight assets; y_1, passenger-kilometres; y_2, non-passenger operating revenue; DEA, efficiency scores estimated under the assumption of constant returns to scale; Sup-EF, super-efficiency; Cross-EF, cross-efficiency scores; NCross-EF, neutral cross-efficiency scores as reported in Table 7 of Wang et al. (2017); Novel-CSW, efficiency scores estimated by a novel CSW approach as reported in TAble 7 of Wang et al. (2017); NCSW, efficiency scores estimated by a normalized weighting method of CSW as reported in Table 7 of Wang et al. (2017); CSW–RA–DEA, efficiency scores estimated by our approach. The ranks of the DMUs are presented inside the brackets

Table 8 Spearman's ranking correlations among the estimated CSW efficiency scores for 14 airlines

	DEA	Sup-EF	CSW–RA–DEA	Cross-EF	NCross-EF	Novel-CSW	NCSW
DEA	**1.000**						
	(0.000)						
Sup-EF	0.936	**1.000**					
	(0.000)	(0.000)					
CSW–RA–DEA	0.800	0.802	**1.000**				
	(0.001)	(0.001)	(0.000)				
Cross-EF	0.857	0.873	0.886	**1.000**			
	(0.000)	(0.000)	(0.000)	(0.000)			
NCross-EF	0.866	0.890	0.886	0.974	**1.000**		
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		
Novel-CSW	0.829	0.829	0.877	0.978	0.952	**1.000**	
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	
NCSW	0.867	0.880	0.918	0.986	0.957	0.977	**1.000**
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)

p-values are presented inside the brackets

References

Acemoglu, D., & Cao, D. (2015). Innovation by entrants and incumbents. *Journal of Economic Theory, 157*, 255–294.

Adler, N., Friedman, L., & Sinuany-Stern, Z. (2002). Review of ranking methods in the data envelopment analysis context. *European Journal of Operational Research, 140*(2), 249–265.

Amiti, M., & Konings, J. (2007). Trade liberalization, intermediate inputs, and productivity: Evidence from Indonesia. *American Economic Review, 97*(5), 1611–1638.

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science, 39*(10), 1261–1264.

Anh, D. L. T., & Gan, C. (2020). Profitability and marketability efficiencies of Vietnam manufacturing firms. *International Journal of Social Economics, 47*(1), 54–71.

Ayyagari, M., Beck, T., & Demirguc-Kunt, A. (2003). *Small and medium enterprises across the globe: A new database*. World Bank.

Bačić, K., Rašić Bakarić, I., & Slijepčević, S. (2018). Sources of productivity differentials in manufacturing in post-transition urban South–East Europe. *Post-Communist Economies, 30*(4), 526–548.

Baldwin, J. R., & Gu, W. (2004). Trade liberalization: Export-market participation, productivity growth, and innovation. *Oxford Review of Economic Policy, 20*(3), 372–392.

Banker, R. D. (1984). Estimating most productive scale size using data envelopment analysis. *European Journal of Operational Research, 17*(1), 35–44.

Belhadi, A., Kamble, S. S., Mani, V., Benkhati, I., & Touriki, F. E. (2021). An ensemble machine learning approach for forecasting credit risk of agricultural SMEs' investments in agriculture 4.0 through supply chain finance. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04366-9

Borchardt, M., Jabbour, C. J. C., de Figueiredo Belém, J., Mani, V., Pereira, G. M., & Ritter, Á. M. (2021). Germinating seeds in dry soil: examining the process of frugal innovation in micro- and small-enterprises at the base of the pyramid. *European Business Review, 34*(3), 297–320.

Boubaker, S., Do, D. T., Hammami, H., & Ly, K. C. (2020). The role of bank affiliation in bank efficiency: a fuzzy multi-objective data envelopment analysis approach. *Annals of Operations Research, 311*, 611–639.

Boubaker, S., Houcine, A., Ftiti, Z., & Masri, H. (2018). Does audit quality affect firms' investment efficiency? *Journal of the Operational Research Society, 69*(10), 1688–1699.

Boubaker, S., Manita, R., & Rouatbi, W. (2019). Large shareholders, control contestability and firm productive efficiency. *Annals of Operations Research, 296*, 591–614.

Breiman, L. (1996). Bagging predictors. *Machine Learning, 24*(2), 123–140.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees* (1st ed.). Routledge.

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. G., & Zelenyuk, V. (2021). LASSO+DEA for small and big wide data. *Omega, 102*, 102419.

CIEM. (2016). *Characteristics of the Vietnamese business environment: Evidence from a SME survey in 2015*. Central Institute for Economic Management (CIEM).

Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., & Roux, S. (2012). The productivity advantages of large cities: Distinguishing agglomeration from firm selection. *Econometrica, 80*(6), 2543–2594.

Contreras, I. (2020). A review of the literature on DEA models under common set of weights. *Journal of Modelling in Management, 15*(4), 1277–1300.

Dao, T. T. T., Mai, X. T. T., Ngo, T., Le, T., & Ho, H. (2021). From efficiency analyses to policy implications: A multilevel hierarchical linear model approach. *International Journal of the Economics of Business, 28*(3), 457–470.

Daraio, C., Simar, L., & Wilson, P. W. (2010). Testing whether two-stage estimation is meaningful in non-parametric models of production. ISBA Discussion Paper.

Davtalab-Olyaie, M. (2019). A secondary goal in DEA cross-efficiency evaluation: A "one home run is much better than two doubles" criterion. *Journal of the Operational Research Society, 70*(5), 807–816.

Greene, W. H. (2008). The econometric approach to efficiency measurement. In H. O. Fried, C. A. K. Lovell, & P. Schmidt (Eds.), *The measurement of productive efficiency and productivity growth* (pp. 92–250). Oxford University Press.

GSO. (2016). *Business results of Vietnamese enterprises in the period 2010–2014*. General Statistics Office (GSO).

Hailu, K. B., & Tanaka, M. (2015). A "true" random effects stochastic frontier analysis for technical efficiency and heterogeneity: Evidence from manufacturing firms in Ethiopia. *Economic Modelling, 50*, 179–192.

Hammami, H., Ngo, T., Tripe, D., & Vo, D.-T. (2020) Ranking with a Euclidean common set of weights in data envelopment analysis: with application to the Eurozone banking sector (Online first). Annals of Operations Research.

Ho, T. H., Nguyen, D. T., Ngo, T., & Le, T. D. (2021). Efficiency in Vietnamese banking: A meta-regression analysis approach. *International Journal of Financial Studies, 9*(3), 41.

Huang, C.-H., & Yang, C.-H. (2016). Ownership, trade, and productivity in Vietnam's manufacturing firms. *Asia-Pacific Journal of Accounting & Economics, 23*(3), 356–371.

IFC. (2012). *IFC and small and medium enterprises*. International Finance Corporation.

Jahanshahloo, G. R., Lotfi, F. H., Sanei, M., & Jelodar, M. F. (2008). Review of ranking models in data envelopment analysis. *Applied Mathematical Sciences, 2*(29), 1431–1448.

Kamble, S. S., Gunasekaran, A., Ghadge, A., & Raut, R. (2020). A performance measurement system for industry 4.0 enabled smart manufacturing system in SMMEs: A review and empirical investigation. *International Journal of Production Economics, 229*, 107853.

Kao, C., & Hung, H.-T. (2005). Data envelopment analysis with common weights: The compromise solution approach. *Journal of Operational Research Society, 56*, 1196–1203.

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

Le, T. D. Q., Ho, T. H., Nguyen, D. T., & Ngo, T. (2021). Fintech credit and bank efficiency: International evidence. *International Journal of Financial Studies, 9*(3), 44.

Le, V., Vu, X.-B., & Nghiem, S. (2018). Technical efficiency of small and medium manufacturing firms in Vietnam: A stochastic meta-frontier analysis. *Economic Analysis and Policy, 59*, 84–91.

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

Manimuthu, A., Venkatesh, V. G., Sreedharan, V. R., & Mani, V. (2021). Modelling and analysis of artificial intelligence for commercial vehicle assembly process in VUCA world: A case study. *International Journal of Production Research, 60*, 4529–4547.

Misiunas, N., Oztekin, A., Chen, Y., & 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.

Nandy, A., & Singh, P. K. (2021). Application of fuzzy DEA and machine learning algorithms in efficiency estimation of paddy producers of rural Eastern India. *Benchmarking: An International Journal, 28*(1), 229–248.

Ngo, T., Le, T., Tran, S. H., Nguyen, A., & Nguyen, C. (2019a). Sources of the performance of manufacturing firms: Evidence from Vietnam. *Post-Communist Economies, 31*(6), 790–804.

Ngo, T., & Tsui, K. W. H. (2021). Estimating the confidence intervals for DEA efficiency scores of Asia–Pacific airlines. *Operational Research, 22*, 3411–3434.

Ngo, T., Vu, H. V., Ho, H., Dao, T. T. T., & Nguyen, H. T. H. (2019b). Performance of fish farms in Vietnam-does financial access help improve their cost efficiency? *International Journal of Financial Studies, 7*(3), 45.

Nguyen, H.-D., Ngo, T., Le, T., Ho, H., & Nguyen, H. T. (2019). The Role of knowledge in sustainable agriculture: Evidence from rice farms' technical efficiency in Hanoi, Vietnam. *Sustainability, 11*(9), 2472.

OECD. (2021). *SME and entrepreneurship policy in Viet Nam*.

Ouenniche, J., & Carrales, S. (2018). Assessing efficiency profiles of UK commercial banks: A DEA analysis with regression-based feedback. *Annals of Operations Research, 266*(1), 551–587.

Pellegrino, G. (2018). Barriers to innovation in young and mature firms. *Journal of Evolutionary Economics, 28*(1), 181–206.

Pham, H. T., Dao, T. L., & Reilly, B. (2010). Technical efficiency in the Vietnamese manufacturing sector. *Journal of International Development, 22*(4), 503–520.

Pilar, P.-G., Marta, A.-P., & Antonio, A. (2018). Profit efficiency and its determinants in small and medium-sized enterprises in Spain. *BRQ Business Research Quarterly, 21*, 238–250.

Rand, J., & Tarp, F. (2020). *Micro, small, and medium enterprises in Vietnam*. Oxford University Press.

Sahoo, P. K., Le, V., & Rath, B. N. (2021). The determinants of firm competitiveness: Evidence from the Indian manufacturing sector. *International Journal of the Economics of Business, 29*, 139–159.

Sariyer, G., Mangla, S. K., Kazancoglu, Y., Tasar, C. O., & Luthra, S. (2021). Data analytics for quality management in Industry 4.0 from a MSME perspective. *Annals of Operations Research*. https://doi.org/10.1007/s10479-021-04215-9

Schoenherr, T., & Speier-Pero, C. (2015). Data science, predictive analytics, and big data in supply chain management: Current state and future potential. *Journal of Business Logistics, 36*(1), 120–132.

Sexton, T. R., Silkman, R. H., & Hogan, A. J. (1986). Data envelopment analysis: Critique and extensions. *New Directions for Program Evaluation, 1986*(32), 73–105.

Smola, A. J., & Schölkopf, B. (2004). A tutorial on support vector regression. *Statistics and Computing, 14*(3), 199–222.

Thaker, K., Charles, V., Pant, A., & Gherman, T. (2021). A DEA and random forest regression approach to studying bank efficiency and corporate governance. *Journal of the Operational Research Society, 73*, 1258–1277.

Thanassoulis, E. (1993). A comparison of regression analysis and data envelopment analysis as alternative methods for performance assessments. *Journal of the Operational Research Society, 44*(11), 1129–1144.

Tone, K., & Tsutsui, M. (2009). Tuning regression results for use in multi-stage data adjustment approach of DEA. *Journal of the Operations Research Society of Japan, 52*(2), 76.

Tran, D. H., & Ngo, D. T. (2014). Performance of the Vietnamese automobile industry: A measurement using DEA. *Asian Journal of Business and Management, 2*(3), 184–191.

Tsai, C.-F., & Chen, M.-L. (2010). Credit rating by hybrid machine learning techniques. *Applied Soft Computing, 10*(2), 374–380.

VCCI, & USAID. (2022). *The Vietnam provincial competitiveness index 2021: Measuring economic governance for business development*. Vietnam Chamber of Commerce and Industry (VCCI) & United States Agency for International Development in Vietnam (USAID), Hà Nội.

Verschelde, M., Dumont, M., Rayp, G., & Merlevede, B. (2016). Semiparametric stochastic metafrontier efficiency of European manufacturing firms. *Journal of Productivity Analysis, 45*(1), 53–69.

Vidal-García, J., Vidal, M., Boubaker, S., & Hassan, M. (2018). The efficiency of mutual funds. *Annals of Operations Research, 267*(1), 555–584.

Vu, H. V., Holmes, M., Tran, T. Q., & Lim, S. (2016). Firm exporting and productivity: What if productivity is no longer a black box. *Baltic Journal of Economics, 16*(2), 95–113.

Wamba, S. F., Gunasekaran, A., Akter, S., Ren, S.J.-F., Dubey, R., & Childe, S. J. (2017). Big data analytics and firm performance: Effects of dynamic capabilities. *Journal of Business Research, 70*, 356–365.

Wang, Q., Liu, Z., & Zhang, Y. (2017). A novel weighting method for finding common weights in DEA. *Asia-Pacific Journal of Operational Research, 34*(05), 1750027.

Wang, Q., Wei, K., Zhang, Y., & Wang, X. (2021). Data envelopment analysis method based on a common set of normalized weights using bargaining game thought. *Computers & Industrial Engineering, 154*, 107047.

Wang, Y.-M., & Chin, K.-S. (2010). Some alternative models for DEA cross-efficiency evaluation. *International Journal of Production Economics, 128*(1), 332–338.

Wang, Y.-M., Luo, Y., & Lan, Y.-X. (2011). Common weights for fully ranking decision making units by regression analysis. *Expert Systems with Applications, 38*(8), 9122–9128.

Wu, D., Yang, Z., & Liang, L. (2006). Using DEA-neural network approach to evaluate branch efficiency of a large Canadian bank. *Expert Systems with Applications, 31*(1), 108–115.

Yang, J.-C. (2006). The efficiency of SMEs in the global market: Measuring the Korean performance. *Journal of Policy Modeling, 28*(8), 861–876.

Zhu, J. (2020). DEA under big data: Data enabled analytics and network data envelopment analysis. *Annals of Operations Research, 309*, 761–783. https://doi.org/10.1007/s10479-020-03668-8

Zhu, N., Zhu, C., & Emrouznejad, A. (2021). A combined machine learning algorithms and DEA method for measuring and predicting the efficiency of Chinese manufacturing listed companies. *Journal of Management Science and Engineering, 6*, 435–448.