Ann Oper Res (2017) 254:235–250
DOI 10.1007/s10479-017-2431-5

ORIGINAL PAPER

An out-of-sample evaluation framework for DEA with application in bankruptcy prediction

Jamal Ouenniche[1] · **Kaoru Tone**[2]

Published online: 17 February 2017

Abstract Nowadays, data envelopment analysis (DEA) is a well-established non-parametric methodology for performance evaluation and benchmarking. DEA has witnessed a widespread use in many application areas since the publication of the seminal paper by Charnes, Cooper and Rhodes in 1978. However, to the best of our knowledge, no published work formally addressed out-of-sample evaluation in DEA. In this paper, we fill this gap by proposing a framework for the out-of-sample evaluation of decision making units. We tested the performance of the proposed framework in risk assessment and bankruptcy prediction of companies listed on the London Stock Exchange. Numerical results demonstrate that the proposed out-of-sample evaluation framework for DEA is capable of delivering an outstanding performance and thus opens a new avenue for research and applications in risk modelling and analysis using DEA as a non-parametric frontier-based classifier and makes DEA a real contender in industry applications in banking and investment.

Keywords Data envelopment analysis · Out-of-sample evaluation · K-Nearest neighbor · Bankruptcy prediction · Risk assessment

1 Introduction

Since the publication of the seminal paper by Charnes, Cooper and Rhodes in 1978, Data envelopment analysis (DEA) has become a well-established non-parametric methodology for performance evaluation and benchmarking. DEA has witnessed a widespread use in many application areas—see Liu et al. (2013) for a recent survey, and Mousavi et al. (2015) and Xu and Ouenniche (2011, 2012a, b) for a recent application area—along with many methodological contributions—see, for example, Banker et al. (1984), Andersen and Petersen

✉ Jamal Ouenniche
Jamal.Ouenniche@ed.ac.uk

1 University of Edinburgh, Business School, 29 Buccleuch Place, Edinburgh EH8 9JS, UK

2 National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan

(1993), Tone (2001, 2002) and Seiford and Zhu (2003). Despite the growing use of DEA, to the best of our knowledge, no published work formally addressed out-of-sample evaluation in DEA. In this paper, we fill this gap by proposing a framework for the out-of-sample evaluation of decision making units.

We illustrate the use of the proposed framework in bankruptcy prediction of companies listed on the London Stock Exchange. Note that prediction of risk class or bankruptcy is one of the major activities in auditing firms' risks and uncertainties. The design of reliable models to predict bankruptcy is crucial for many decision making processes. Bankruptcy prediction models could be divided into two broad categories depending on whether they are static (see, for example, Altman 1968, 1983; Taffler 1984; Theodossiou 1991; Ohlson 1980; Zmijewski 1984) or dynamic (see, for example, Shumway 2001; Bharath and Shumway 2008; Hillegeist et al. 2004). In this paper we shall focus on the first category of models to illustrate how out-of-sample evaluation of companies could be performed. The most popular static bankruptcy prediction models are based on statistical methodologies (e.g., Altman 1968, 1983; Taffler 1984), stochastic methodologies (e.g., Theodossiou 1991; Ohlson 1980; Zmijewski 1984), and artificial intelligence methodologies (e.g., Kim and Han 2003; Li and Sun 2011; Zhang et al. 1999; Shin et al. 2005). DEA methodologies are increasingly gaining popularity in bankruptcy prediction (e.g., Cielen et al. 2004; Paradi et al. 2004; Premachandra et al. 2011; Shetty et al. 2012). However, the issue of out-of-sample evaluation remains to be addressed when DEA is used as a classifier.

The remainder of this paper is organised as follows. In Sect. 2, we propose a formal framework for performing out-of-sample evaluation in DEA. In Sect. 3, we provide information on the bankruptcy data we used along with details on the design of our experiment, and present our empirical findings. Finally, Sect. 4 concludes the paper.

2 A framework for out-of-sample evaluation in DEA

Nowadays, out-of-sample evaluation of statistical, stochastic and artificial intelligence methodologies for prediction of both continuous and discrete variables is commonly used for validating prediction models and testing their performance before actual implementation. The rationale for using out-of-sample testing lies in the following well known facts. First, models or methods selected based on in-sample performance may not best predict post-sample data. Second, in-sample errors are likely to understate prediction errors. Third, for continuous variables, prediction intervals built on in-sample standard errors are likely to be too narrow. The setup of the standard out-of-sample analysis framework requires one to split the historical data set into two subsets, where the first subset often referred to as a training set, an estimation set, or an initialization set is used to estimate the parameters of a model, whereas the second subset generally referred to as the test set or the handout set is used to test the prediction performance of the fitted model. The counterpart of this testing framework is lacking in DEA. In this paper, we propose an out-of-sample evaluation framework for static DEA models. The proposed framework in general in that it can be used for any classification problem or number of classes and any application. Note that, without loss of generality, the proposed framework is customized for a bankruptcy prediction application with two risk classes (e.g., bankrupt class and non-bankrupt class, or low risk of bankruptcy class and high risk of bankruptcy class), as customary in most research on bankruptcy prediction, for the sake of illustrating the empirical performance of our framework. Obviously this risk classification into two categories or classes could be refined, if the researcher/analyst wished

to do so, into more than two classes when the presence of non-zero slacks is suspected or proven to be a driver of bankruptcy; for example, one might be interested in refining each of the above mentioned risk classes into two subclasses depending on whether the slacks of a bankrupt (respectively, non-bankrupt) DMU sum to zero or not. In other practical settings, the researcher/analyst might be interested in the level or degree of distress prior to bankruptcy in which case one might also consider more than two risk or distress classes. In the remaining of this paper, we denote the variable on risk class belonging as Y. Hereafter, we describe the main steps of the proposed out-of-sample evaluation framework for DEA:

Input: data set of historical observations, say X, where each observation is a DMU (e.g., firm-year observations where firms are listed on the London Stock Exchange) along with the corresponding available information (e.g., financial ratios) and the observed risk or bankruptcy status Y;

Step 1: divide the "historical" sample X into an estimation set X_E – hereafter referred to as training sample I—and a test set X_T – hereafter referred to as test sample I. Then, customize X_E and X_T for the implementation of a specific DEA model by only retaining the input and output information used by the DEA model, which results in X_E^{I-O} and X_T^{I-O} – hereafter referred to as training sample II and test sample II, respectively;

Step 2: solve an appropriate DEA model to compute DEA efficiency scores and the slacks for DMUs in training sample X_E^{I-O} and classify them according to a user-specified classification rule into, for example, risk or bankruptcy classes, say $\hat{Y}_E^{I-O}$. Then, compare the DEA based classification of DMUs in X_E^{I-O} into risk classes; that is, the predicted risk classes $\hat{Y}_E^{I-O}$, with the observed risk classes Y_E of DMUs in the training sample, and compute the relevant in-sample performance statistics;

Step 3: use an appropriate algorithm to classify DMUs in X_T^{I-O} into, for example, risk or bankruptcy classes, say $\hat{Y}_T^{I-O}$. Then, compare the predicted risk classes $\hat{Y}_T^{I-O}$ with the observed risk classes Y_T and compute the relevant out-of-sample performance statistics;

Step 4: for each DMU in X_T^{I-O}, use the multiplier(s) of their closest DMU(s) in X_E^{I-O} to compute its efficiency score, if required.

Output: in-sample and out-of-sample classifications or risk class belongings of DMUs along with the corresponding performance statistics, and DEA efficiency scores of DMUs in both training and test samples.

Note that this procedure is generic in nature. The flowchart of the proposed framework is depicted in Fig. 1 to provide a snapshot of its design. The implementation of this generic procedure requires several decisions to be made. First, one has to decide on which DEA model to use for assessing the efficiency of DMUs in X_E^{I-O}. Second, one has to decide on which decision rule to use for classifying DMUs in X_E^{I-O}. Third, one has to decide on which algorithm to use for classifying DMUs in X_T^{I-O}. Finally, one has to decide on how to exploit the information on the performance of similar DMUs in X_E^{I-O} to assess the performance of DMUs in X_T^{I-O}. Hereafter, we shall present how we choose to address these questions.

2.1 DEA model for assessing the efficiency of DMUs in the training sample

A variety of DEA models could be used for this task. However, the final choice depends on the type of application one is dealing with and the suitability of the DEA model or analysis for

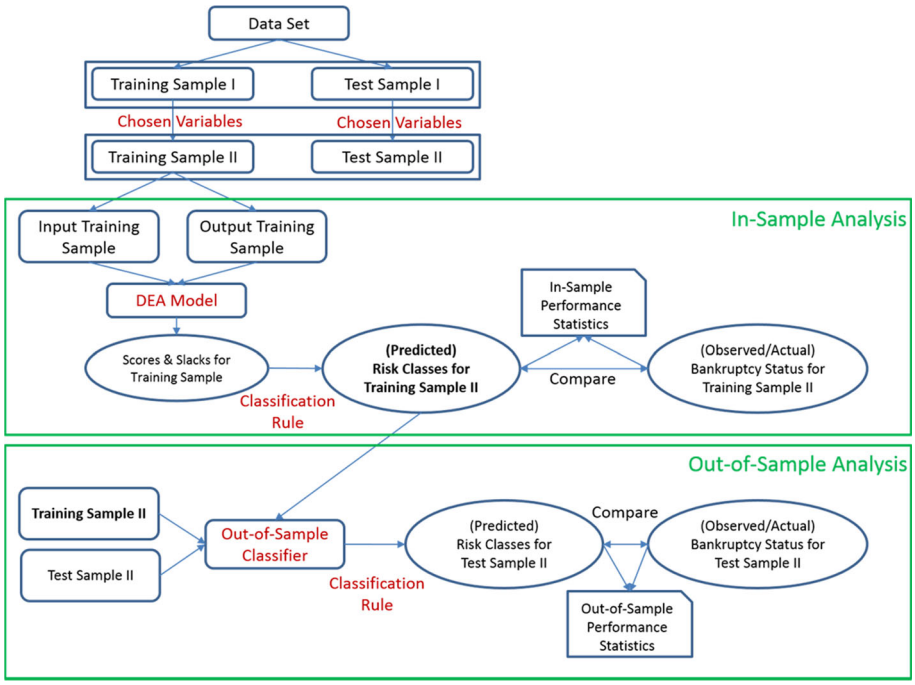

Fig. 1 Flowchart of out-of-sample evaluation framework for static DEA models

such application. For the bankruptcy application, two main categories of DEA models could be used; namely, best efficiency frontier-based models (e.g., Charnes et al. 1978; Banker et al. 1984; Tone 2001) and worst efficiency frontier-based models (e.g., Paradi et al. 2004). Within each of these categories one could choose from a variety of DEA models. Note that the main difference between the best efficiency frontier-based models and the worst efficiency frontier-based models lies in the choice of the definition of the efficiency frontier. To be more specific, best efficiency frontier-based DEA models assume that the efficiency frontier is made of the best performers, whereas the worst efficiency frontier-based DEA models assume that the efficiency frontier is made of the worst performers (i.e., riskiest DMUs). In risk modelling and analysis applications, such as bankruptcy prediction, both types of frontiers or DEA models are appropriate to use; however, the classification rules used in step 2 and step 3 of the detailed procedure would have to be chosen accordingly.

For illustration purposes, in our empirical investigation, we used both a BCC model (Banker et al. 1984) and an SBM model (Tone 2001) and implemented each of them within the best efficiency frontier framework. Notice that, since our data consists of financial ratios which could take negative values, the SBM model was implemented within a variable return-to-scale framework; that is, the convexity constraint was added to the model. These models are presented in Tables 1, 2, where the parameter $x_{i,j}$ denote the amount of input i used by DMU_j, the parameter $y_{r,j}$ denote the amount of output r produced by DMU_j, the decision variable λ_j denote the weight assigned to DMU_j's inputs and outputs in constructing the ideal benchmark of a given DMU, say DMU_k, the decision variable θ_k denote the technical efficiency score of DMU_k, and the decision variable ρ_k denote the slacks-based measure (SBM) for DMU_k.

Table 1 Best efficiency frontier BCC models

Formulation	Description
θ_k	Objective; that is, technical efficiency score. This objective is to be minimized in the input-oriented version of the model and maximized in the output-oriented version of the model
$\sum_{j=1}^{n} \lambda_j x_{i,j} \leq \theta_k \cdot x_{i,k};\ \forall i$ or $\sum_{j=1}^{n} \lambda_j x_{i,j} \leq x_{i,k};\ \forall i$	For each input i ($i = 1, \ldots, m$), the amount used by DMU_k's "ideal" benchmark; i.e., its projection on the efficient frontier ($\sum_{j=1}^{n} \lambda_j x_{i,j}$), should at most be equal to the amount used by DMU_k whether revised (i.e., amount of input i adjusted for the degree of technical efficiency of DMU_k) or not depending on whether the model is input-oriented or not
$\sum_{j=1}^{n} \lambda_j y_{r,j} \geq y_{r,k};\ \forall r$ or $\sum_{j=1}^{n} \lambda_j y_{r,j} \geq \theta_k \cdot y_{r,k};\ \forall r$	For each output r ($r = 1, \ldots, s$), the amount produced by DMU_k's "ideal" benchmark; i.e., its projection on the efficient frontier ($\sum_{j=1}^{n} \lambda_j y_{r,j}$), should be at least as large as the amount produced by DMU_k whether revised (i.e., amount of output r adjusted for the degree of technical efficiency of DMU_k) or not depending on whether the model is output-oriented or not
$\sum_{j=1}^{n} \lambda_j = 1$	The technology is required to be convex
$\lambda_j \geq 0;\ \forall j$	Non-negativity requirements

Table 2 Best efficiency frontier SBM models

Formulation	Description
$\rho_k = 1 - \frac{1}{m}\left(\sum_{i=1}^{m} \frac{s_{i,k}^{-}}{x_{i,k}}\right)$	Objective; that is, input-oriented SBM measure
$\rho_k = \frac{1}{1+\frac{1}{s}\left(\sum_{r=1}^{s} \frac{s_{r,k}^{+}}{y_{r,k}}\right)}$	Objective; that is, output-oriented SBM measure
$\rho_k = \frac{1-\frac{1}{m}\left(\sum_{i=1}^{m} \frac{s_{i,k}^{-}}{x_{i,k}}\right)}{1+\frac{1}{s}\left(\sum_{r=1}^{s} \frac{s_{r,k}^{+}}{y_{r,k}}\right)}$	Objective; that is, Non-Oriented SBM measure
$\sum_{j=1}^{n} \lambda_j x_{i,j} + s_{i,k}^{-} = x_{i,k};\ \forall i$	For each input i ($i = 1, \ldots, m$), the amount used by DMU_k's "ideal" benchmark; i.e., its projection on the efficient frontier, should at most be equal to the amount used by DMU_k; that is: $\sum_{j=1}^{n} \lambda_j x_{i,j} \leq x_{i,k};\ \forall i$
$\sum_{j=1}^{n} \lambda_j y_{r,j} - s_{r,k}^{+} = y_{r,k};\ \forall r$	For each output r ($r = 1, \ldots, s$), the amount produced by DMU_k's "ideal" benchmark; i.e., its projection on the efficient frontier, should be at least as large as the amount produced by DMU_k; that is: $\sum_{j=1}^{n} \lambda_j y_{r,j} \geq y_{r,k};\ \forall r$
$\sum_{j=1}^{n} \lambda_j = 1$	The technology is required to be convex
$\lambda_j \geq 0;\ \forall j;\ s_{i,k}^{-};\ \forall i;\ s_{r,k}^{+};\ \forall r$	Non-negativity requirements

Table 3 Generic procedure for computing an optimal DEA score-based cut-off point and the corresponding classification

Input: choice of a performance measure π and a non-linear programming search algorithm according to the properties of π
Step 1: compute ξ_{LB} and ξ_{UB}
Step 2: find the optimal value of ξ with respect to π, say ξ^*, within the interval $[\xi_{LB}, \xi_{UB}]$ using the chosen non-linear programming search algorithm
Step 3: classify DMUs in X_E^{I-O} into two classes; namely bankrupt and non-bankrupt firms or DMUs; that is, determine $\hat{Y}_E^{I-O}$ so that DMUs with DEA scores less (respectively, greater) than ξ^* are assigned to a bankruptcy class and those with DEA scores greater (respectively, less) than or equal to ξ^* are assigned to a non-bankruptcy class if a best practice (respectively, worse practice) efficiency frontier framework was adopted to compute DEA scores
Output: optimal DEA score-based cut-off point ξ^* along with the predicted risk classes $\hat{Y}_E^{I-O}$

2.2 Decision rule for classifying DMUs in the training sample

Several decision rules could be used to classify the DMUs in the training sample. Obviously the choice of a decision rule for classification depends on the nature of the classification problem. To be more specific, decision rules would vary depending on whether one is concerned with a two-class problem or a multi-class problem. In bankruptcy prediction we are concerned with a two-class problem; therefore, we shall provide a solution that is suitable for these problems. In fact, we propose a DEA score-based cut-off point procedure to classify DMUs in X_E^{I-O}. The proposed procedure involves solving an optimization problem whereby the DEA score-based cut-off point, say ξ, is determined so as to optimize a given performance measure, say π, over an interval with a lower bound, say ξ_{LB}, equal to the smallest DEA score of DMUs in X_E^{I-O} and an upper bound, say ξ_{UB}, equal to the largest DEA score of DMUs in X_E^{I-O}. In sum, the proposed procedure is based on a performance measure-dependent approach—see Table 3 for a generic procedure. Note that, in most applications, the performance measure π is a non-linear function. The choice of a specific optimization algorithm for the implementation of the generic procedure outlined in Table 3 depends on whether the performance measure π is differentiable or not and if it is non-differentiable, whether it is quasiconvex or not. To be more specific, if π is differentiable, then one could choose Bisection Search; if π is twice differentiable, then one could choose Newton's Method; if π is non-differentiable but quasiconvex, then one could choose Golden Section Search, Fibonacci Search, Dichotomous Search, or a brute force search such as Uniform Search. For details on these standard non-linear programming algorithms, the reader is referred to the excellent book on non-linear programming by Bazaraa et al. (2006). Notice that the last step of this generic procedure classifies DMUs in the training sample into two classes; namely bankrupt and non-bankrupt firms or DMUs, and thus the output is the optimal DEA score-based cut-off point ξ along with the predicted risk classes $\hat{Y}_E^{I-O}$.

2.3 Algorithm for classifying DMUs in the test sample

A variety of algorithms could be used for out-of-sample classification of DMUs in X_T^{I-O} ranging from standard statistical and stochastic methodologies to artificial intelligence methodologies. In this paper, we propose an instance of our generic out-of-sample evaluation procedure for DEA where the out-of-sample classification of DMUs in X_T^{I-O} is performed

Initialization Step
Choose the Case Base as X_E^{I-O} and the Query Set as X_T^{I-O};
Choose a distance metric d;
Choose a classification criterion;
Iterative Step
// Compute distances between queries and cases
FOR $i = 1$ to $|X_T^{I-O}|$ {
  FOR $j = 1$ to $|X_E^{I-O}|$ {
    Compute $d(DMU_i, DMU_j)$; }}
// Sort cases in ascending order of their distances to queries and classify queries
FOR $i = 1$ to $|X_T^{I-O}|$ {
  Sort the list $L_i = \left\{\left(j, d(DMU_i, DMU_j)\right);\ j = 1, \ldots |X_E^{I-O}|\right\}$ in ascending order of distances and use the first k entries in the list $L_i(1{:}k, .)$ to classify DMU_i according to the chosen criterion;}

Fig. 2 Pseud-code of the k-NN algorithm

using a k-Nearest Neighbor (k-NN) algorithm, which itself is an instance of case-based reasoning. The pseudo-code for k-NN is customized for our application and is summarized in Fig. 2. Note that the k-NN algorithm is also generic in that a number of implementation decisions have to be made; namely, the size of the neighborhood k, the similarity or distance metric, and the classification criterion. In our experiments, we tested several values of k as well as several distance metrics (i.e., Euclidean, Standardized Euclidean, Cityblock, Hamming, Jaccard, Cosine, Correlation, Mahalanobis). As to the classification criterion, we opted for the most commonly used one; that is, majority vote. Note that, when computing the distance between two DMUs, each DMU is represented by its vector of inputs and outputs.

2.4 Computing efficiency scores of DMUs in the test sample

In order to compute the DEA score of those DMUs in X_T^{I-O}, one could opt for one of three possible approaches. First, one could simply solve a DEA model for each DMU in X_T^{I-O} – although this option is a valid one, it would defeat the purpose of out-of-sample evaluation. Instead, we propose to either use the multipliers of a most similar or closest DMU in X_E^{I-O} to compute the DEA score of a DMU in X_T^{I-O}, or use the multipliers of the ℓ ($\ell > 1$) most similar or closest DMUs in X_E^{I-O} to compute ℓ DEA scores of a DMU in X_T^{I-O} and take their average or weighted average as the final score.

To conclude this section, we would like to provide some explanation as to why the proposed framework should produce good results. As the reader is aware of by now, the proposed out-of-sample evaluation framework is based on an instance of the case-based reasoning (CBR) methodology; namely, k-NN algorithm. CBR is a generic problem solving methodology, which solves a specific problem by exploiting solutions to similar problems. In sum, CBR relies on past experience and comparison to the current experience and therefore uses analogy by similarity. To be more specific, the basic methodological process of this artificial intelligence methodology involves pattern matching and classification. In our bankruptcy application, pattern matching would serve to identify DMUs with similar risk profiles (e.g., liquidity profiles in our experiments) and therefore is well equipped to discriminate between bankrupt and non-bankrupt firms. The extent of its empirical performance however would depend on whether the data or case base is noisy or not, the choice of the similarity criteria and their measures, the relevance of the features selected (i.e., inputs and outputs in the DEA

context) and their weights, if any, and the choice of the classification rule, also known as a target function, as well as the quality of approximation of the target function. In our case, k-NN serves as a local approximation. For more details on CBR, the reader is referred to, for example, Richter and Weber (2013).

In the next section, we shall test the performance of our out-of-sample evaluation framework for DEA and report our numerical results.

3 Empirical analysis

In this section, we first describe the process of data gathering and sample selection (see Sect. 3.1). Then, we present the design of our experiment (see Sect. 3.2). Finally, we present and discuss our numerical results (see Sect. 3.3).

3.1 Data and sample selection

In this paper, we first considered all UK firms listed on the London Stock Exchange (LSE) during a 5 years period from 2010 through 2014 and defined the bankrupt firms using the London Share Price Database (LSPD) codes 16 (i.e., firm has receiver appointed or is in liquidation), 20 (i.e., firm is in administration or administrative receivership), and 21 (i.e., firm is cancelled and assumed valueless); the remaining firms are classified as non-bankrupt. Then, we further reduced such dataset by excluding both financial and utilities firms, on one hand, and those firms with less than 5 months lag between the reporting date and the fiscal year, on the other hand. As a result of using these data reduction rules, the final dataset consists of 6605 firm-year observations including 407 (6.16%) observations related to bankrupt firms and 6198 (94.38%) observations related to non-bankrupt firms. Therefore, we have a total of 6605 decision making units (DMUs). As to the selection of the training sample and the test sample, we have chosen the size of the training sample to be twice the size of the test sample; that is, 2/3 of the total number of DMUs were used in the training sample and the remaining 1/3 were used in the test sample. The selection of observations was done with random sampling without replacement so as to ensure that both the training sample and the test sample have the same proportions of bankrupt and non-bankrupt firms. A total of thirty pairs of training sample-test sample were generated.

3.2 Design of experiment

In our experiment, we reworked a standard and well known parametric model in the DEA framework; namely, the multivariate discriminant analysis (MDA) model of Taffler (1984) to provide some empirical evidence on the merit of the proposed out-of-sample evaluation framework for DEA. Recall that Taffler's model makes use of four explanatory variables; namely, current liabilities to total assets, number of credit intervals, profit before tax to current liabilities, and current assets to total liabilities. In our DEA models, current liabilities to total assets and number of credit intervals were used as inputs, whereas profit before tax to current liabilities and current assets to total liabilities were used as outputs. We report on the performance of our out-of-sample evaluation framework for DEA using the commonly used metrics; namely, type I error (T1), type II error (T2), sensitivity (Sen) and specificity (Spe). Recall that T1 is the proportion of bankrupt firms predicted as non-bankrupt; T2 is the proportion of non-bankrupt firms predicted as bankrupt; Sen is the proportion of non-bankrupt firms predicted as non-bankrupt; and Spe is the proportion of bankrupt firms predicted as bankrupt.

Table 4 Summary statistics of in-sample performance of DEA models

Performance measures	T1	T2	Sen	Spe
BCC-IO				
Min	0.0000	0.0000	1.0000	0.9926
Max	0.0074	0.0000	1.0000	1.0000
Average	0.0038	0.0000	1.0000	0.9962
SD	0.0025	0.0000	0.0000	0.0025
BCC-OO				
Min	0.0000	0.0000	1.0000	1.0000
Max	0.0000	0.0000	1.0000	1.0000
Average	0.0000	0.0000	1.0000	1.0000
SD	0.0000	0.0000	0.0000	0.0000
SBM-IO				
Min	0.0000	0.0000	1.0000	0.9926
Max	0.0074	0.0000	1.0000	1.0000
Average	0.0032	0.0000	1.0000	0.9968
SD	0.0021	0.0000	0.0000	0.0021
SBM-OO				
Min	0.0000	0.0000	1.0000	0.9926
Max	0.0074	0.0000	1.0000	1.0000
Average	0.0030	0.0000	1.0000	0.9970
SD	0.0020	0.0000	0.0000	0.0020
SBM				
Min	0.0000	0.0000	1.0000	0.9926
Max	0.0074	0.0000	1.0000	1.0000
Average	0.0030	0.0000	1.0000	0.9970
SD	0.0020	0.0000	0.0000	0.0020

Table 5 Summary statistics of in-sample and out-of-sample performance of MDAs

Performance measures	T1	T2	Sen	Spe
In-sample MDA				
Min	0.9705	0.0019	0.9937	0.0000
Max	1.0000	0.0063	0.9981	0.0295
Average	0.9882	0.0026	0.9974	0.0118
SD	0.0067	0.0009	0.0009	0.0067
Out-of-sample MDA				
Min	0.0000	0.0000	0.0015	0.0000
Max	1.0000	0.9985	1.0000	1.0000
Average	0.8220	0.1701	0.8299	0.1780
SD	0.3743	0.3766	0.3766	0.3743

3.3 Results

Hereafter, we shall provide a summary of our empirical results and findings. Table 4 provides a summary of statistics on the performance of the MDA model of Taffler (1984) reworked

Table 6 Summary statistics of out-of-sample performance of BCC-IO

Metric	Statistics	Performance measures			
		T1	T2	Sen	Spe
Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1005	0.0000	1.0000	0.8995
	SD	0.3017	0.0000	0.0000	0.3017
Standardized Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1206	0.0000	1.0000	0.8794
	SD	0.3127	0.0000	0.0000	0.3127
Cityblock	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1000	0.0000	1.0000	0.9000
	SD	0.3018	0.0000	0.0000	0.3018
Hamming	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Jaccard	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Cosine	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1449	0.0000	1.0000	0.8551
	SD	0.3399	0.0000	0.0000	0.3399
Correlation	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1456	0.0000	1.0000	0.8544
	SD	0.3399	0.0000	0.0000	0.3399
Mahalanobis	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1039	0.0000	1.0000	0.8961
	SD	0.2865	0.0000	0.0000	0.2865

within the best efficiency frontier framework using BCC and SMB models. Note that both in-sample and out-of-sample statistics reported correspond to DEA score-based cut-off points optimized for each performance measure separately (i.e., T1, T2, Sen, Spe). Note also that we run tests for several values of the size of the neighborhood k (i.e., 3, 5, 7); however, the results reported are for $k = 3$ since higher values delivered very close performances but required more computations.

With respect to in-sample performance, our results demonstrate that DEA provides an outstanding classifier regardless of the choices of classification measures and DEA models—

Table 7 Summary statistics of out-of-sample performance of BCC-OO

Metric	Statistics	Performance measures			
		T1	T2	Sen	Spe
Euclidean	Min	0.0000	0.0000	1.0000	0.0074
	Max	0.9926	0.0000	1.0000	1.0000
	Average	0.0990	0.0000	1.0000	0.9010
	SD	0.3021	0.0000	0.0000	0.3021
Standardized Euclidean	Min	0.0000	0.0000	1.0000	0.0074
	Max	0.9926	0.0000	1.0000	1.0000
	Average	0.1650	0.0000	1.0000	0.8350
	SD	0.3738	0.0000	0.0000	0.3738
Cityblock	Min	0.0000	0.0000	1.0000	0.0074
	Max	0.9926	0.0000	1.0000	1.0000
	Average	0.0973	0.0000	1.0000	0.9027
	SD	0.2954	0.0000	0.0000	0.2954
Hamming	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Jaccard	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Cosine	Min	0.0000	0.0000	1.0000	0.0368
	Max	0.9632	0.0000	1.0000	1.0000
	Average	0.0321	0.0000	1.0000	0.9679
	SD	0.1759	0.0000	0.0000	0.1759
Correlation	Min	0.0000	0.0000	1.0000	0.0441
	Max	0.9559	0.0000	1.0000	1.0000
	Average	0.0326	0.0000	1.0000	0.9674
	SD	0.1744	0.0000	0.0000	0.1744
Mahalanobis	Min	0.0000	0.0000	1.0000	0.0074
	Max	0.9926	0.0000	1.0000	1.0000
	Average	0.1885	0.0000	1.0000	0.8115
	SD	0.3856	0.0000	0.0000	0.3856

see Table 4. In fact, in-sample, DEA does not wrongly classify any non-bankrupt firm as demonstrated by type II error of 0% and sensitivity of 100%. On the other hand, most bankrupt firms are properly classified as demonstrated by a very small range (0–0.74%) and a very small average (0.38%) of type I error, and a very small range (99.26–100%) of specificity. However, BCC-OO delivers the ideal performance with T1 and T2 being 0% and sensitivity and specificity being 100%. An additional evidence of the superiority of DEA over Discriminant Analysis in-sample is provided in Table 5 with differences, for example, in average performance of 98% on T1 and Spe and 0.26% on T2 and Sen in favor of DEA.

Table 8 Summary statistics of out-of-sample performance of SBM-IO

Metric	Statistics	Performance measures			
		T1	T2	Sen	Spe
Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0360	0.0000	1.0000	0.9640
	SD	0.1821	0.0000	0.0000	0.1821
Standardized Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0355	0.0000	1.0000	0.9645
	SD	0.1822	0.0000	0.0000	0.1822
Cityblock	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0363	0.0000	1.0000	0.9637
	SD	0.1821	0.0000	0.0000	0.1821
Hamming	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Jaccard	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Cosine	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1770	0.0000	1.0000	0.8230
	SD	0.3731	0.0000	0.0000	0.3731
Correlation	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1775	0.0000	1.0000	0.8225
	SD	0.3732	0.0000	0.0000	0.3732
Mahalanobis	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0355	0.0000	1.0000	0.9645
	SD	0.1822	0.0000	0.0000	0.1822

Next, we provide empirical evidence to demonstrate that the proposed out-of-sample evaluation framework achieved a very high performance in classifying DMUs into the right risk category—see Tables 6, 7, 8, 9 and 10. In fact, regardless of which DEA model is chosen to compute the scores, the out-of-sample performance of the proposed framework is ideal—with T1 and T2 being 0% and sensitivity and specificity being 100%—when Hamming and Jaccard metrics are used to compute the distances between training sample and test sample observations or DMUs. As to the remaining metrics, they deliver average performances ranging from −0.05 to 18%. It is worthy to mention however that the choice of SBM-OO

Table 9 Summary statistics of out-of-sample performance of SBM-OO

Metric	Statistics	Performance measures			
		T1	T2	Sen	Spe
Euclidean	Min	0.0000	0.0000	1.0000	0.9926
	Max	0.0074	0.0000	1.0000	1.0000
	Average	0.0025	0.0000	1.0000	0.9975
	SD	0.0035	0.0000	0.0000	0.0035
Standardized Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0353	0.0000	1.0000	0.9647
	SD	0.1822	0.0000	0.0000	0.1822
Cityblock	Min	0.0000	0.0000	1.0000	0.9853
	Max	0.0147	0.0000	1.0000	1.0000
	Average	0.0025	0.0000	1.0000	0.9975
	SD	0.0040	0.0000	0.0000	0.0040
Hamming	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Jaccard	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Cosine	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1434	0.0000	1.0000	0.8566
	SD	0.3404	0.0000	0.0000	0.3404
Correlation	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1446	0.0000	1.0000	0.8554
	SD	0.3402	0.0000	0.0000	0.3402
Mahalanobis	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0353	0.0000	1.0000	0.9647
	SD	0.1822	0.0000	0.0000	0.1822

and SBM models combined with Euclidean and Cityblock metrics drive the performance of the proposed framework to an unexpected high level with an average performance of −0.05% suggesting that the proposed framework fed with the right decisions could even strengthen in-sample DEA analysis. Once again, the proposed out-of-sample evaluation framework for DEA proves to be superior to Discriminant Analysis out-of-sample (see Table 5) with differences, for example, in average performance of 79–98% on T1, 0.26% on T2 and Sen, and 63–82% on Spe in favor of DEA.

Table 10 Summary statistics of out-of-sample performance of SBM

Metric	Statistics	Performance measures			
		T1	T2	Sen	Spe
Euclidean	Min	0.0000	0.0000	1.0000	0.9926
	Max	0.0074	0.0000	1.0000	1.0000
	Average	0.0025	0.0000	1.0000	0.9975
	SD	0.0035	0.0000	0.0000	0.0035
Standardized Euclidean	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0554	0.0000	1.0000	0.9446
	SD	0.2102	0.0000	0.0000	0.2102
Cityblock	Min	0.0000	0.0000	1.0000	0.9853
	Max	0.0147	0.0000	1.0000	1.0000
	Average	0.0025	0.0000	1.0000	0.9975
	SD	0.0040	0.0000	0.0000	0.0040
Hamming	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Jaccard	Min	0.0000	0.0000	1.0000	1.0000
	Max	0.0000	0.0000	1.0000	1.0000
	Average	0.0000	0.0000	1.0000	1.0000
	SD	0.0000	0.0000	0.0000	0.0000
Cosine	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1441	0.0000	1.0000	0.8559
	SD	0.3401	0.0000	0.0000	0.3401
Correlation	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.1446	0.0000	1.0000	0.8554
	SD	0.3402	0.0000	0.0000	0.3402
Mahalanobis	Min	0.0000	0.0000	1.0000	0.0000
	Max	1.0000	0.0000	1.0000	1.0000
	Average	0.0623	0.0000	1.0000	0.9377
	SD	0.2312	0.0000	0.0000	0.2312

4 Conclusions

Out-of-sample evaluation is commonly used for validating prediction models of both continuous and discrete variables and testing their performance. The counterpart of this evaluation framework is lacking in DEA. This paper fills this gap. In fact, we proposed a generic out-of-sample evaluation framework for DEA and tested the performance of an instance of it in bankruptcy prediction. The accuracy of our framework, as suggested by our numerical results, suggests that this tool could prove valuable in industry implementations of DEA

models in bankruptcy prediction and credit scoring. We also provided empirical evidence that DEA as a classifier is a real contender to Discriminant Analysis, which is one of the most commonly used classifiers by practitioners.

References

Andersen, P., & Petersen, N. C. (1993). A procedure for ranking efficient units in data envelopment analysis. *Management Science*, *39*, 1261–1294.

Altman, E. (1968). Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*, *23*(4), 589–609.

Altman, E. (1983). *Corporate financial distress: A complete guide to predicting, avoiding and dealing with bankruptcy*. Hoboken: Wiley.

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Models for the estimation of technical and scale inefficiencies in data envelopment analysis. *Management Science*, *30*, 1078–1092.

Bazaraa, M. S., Sherali, H. D., & Shetty, C. M. (2006). *Nonlinear programming: Theory and algorithms* (3rd ed.). New Jersey: Wiley.

Bharath, S. T., & Shumway, T. (2008). Forecasting default with the Merton distance to default model. *Review of Financial Studies*, *21*(3), 1339–1369.

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

Cielen, A., Peeters, L., & Vanhoof, K. (2004). Bankruptcy prediction using a data envelopment analysis. *European Journal of Operational Research*, *154*, 526–532.

Hillegeist, S. A., Keating, E. K., Cram, D. P., & Lundstedt, K. G. (2004). Assessing the probability of bankruptcy. *Review of Accounting Studies*, *9*(1), 5–34.

Kim, M.-J., & Han, I. (2003). The discovery of experts' decision rules from qualitative bankruptcy data using genetic algorithms. *Expert Systems with Applications*, *25*(4), 637–646.

Li, H., & Sun, J. (2011). Predicting business failure using forward ranking-order case-based reasoning. *Expert Systems with Applications*, *38*(4), 3075–3084.

Liu, J. S., Lu, L. Y. Y., Lu, W.-W., & Lin, B. J. Y. (2013). A survey of DEA applications. *Omega*, *41*, 893–902.

Mousavi, M. M., Ouenniche, J., & Xu, B. (2015). Performance evaluation of bankruptcy prediction models: An orientation-free super-efficiency dea-based framework. *International Review of Financial Analysis*, *42*, 64–74.

Ohlson, J. A. (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, *18*(1), 109–131.

Paradi, J. C., Asmild, M., & Simak, P. C. (2004). Using DEA and worst practice DEA in credit risk evaluation. *Journal of Productivity Analysis*, *21*, 153–165.

Premachandra, I. M., Chen, Y., & Watson, J. (2011). DEA as a tool for predicting corporate failure and success: A case of bankruptcy assessment. *Omega*, *39*, 620–626.

Richter, M. M., & Weber, R. O. (2013). *Case-based reasoning: A textbook*. Berlin: Springer.

Seiford, L. M., & Zhu, J. (2003). Context-dependent data envelopment analysis—measuring attractiveness and progress. *Omega*, *31*, 397–408.

Shetty, U., Pakkala, T. P. M., & Mallikarjunappa, T. (2012). A modified directional distance formulation of DEA to assess bankruptcy: An application to IT/ITES companies in India. *Expert Systems with Applications*, *39*, 1988–1997.

Shin, K.-S., Lee, T. S., & Kim, H.-J. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, *28*, 127–135.

Shumway, T. (2001). Forecasting bankruptcy more accurately: A simple hazard model. *Journal of Business, 74*(1), 101–124.

Taffler, R. J. (1984). Empirical models for the monitoring of UK corporations. *Journal of Banking & Finance*, *8*(2), 199–227.

Theodossiou, P. (1991). Alternative models for assessing the financial condition of business in Greece. *Journal of Business Finance & Accounting*, *18*(5), 697–720.

Tone, K. (2001). A slacks-based measure of efficiency in data envelopment analysis. *European Journal of Operational Research*, *130*, 498–509.

Tone, K. (2002). A slacks-based measure of super-efficiency in data envelopment analysis. *European Journal of Operational Research*, *143*, 32–41.

Xu, B., & Ouenniche, J. (2012). Performance Evaluation of Competing Forecasting Models - A Multidimensional Framework based on Multi-Criteria Decision Analysis. *Expert Systems with Applications*, *39*(9), 8312–8324.

Xu, B., & Ouenniche, J. (2012). A data envelopment analysis-based framework for the relative performance evaluation of competing crude oil prices' volatility forecasting models. *Energy Economics*, *34*(2), 576–583.

Xu, B., & Ouenniche, J. (2011). A multidimensional framework for performance evaluation of forecasting models: Context-dependent DEA. *Applied Financial Economics*, *21*(24), 1873–1890.

Zhang, G., Hu, M. Y., Patuwo, B. E., & Indro, D. C. (1999). Artificial neural networks in bankruptcy prediction: General framework and cross-validation analysis. *European Journal of Operational Research*, *116*, 16–32.

Zmijewski, M. E. (1984). Methodological issues related to the estimation of financial distress prediction models. *Journal of Accounting Research*, *22*, 59–82.