Expert Systems with Applications 39 (2012) 10495–10499

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Uncover the predictive structure of healthcare efficiency applying a bootstrapped data envelopment analysis

Arianna De Nicola [a], Simone Gitto [b], Paolo Mancuso [c,*]

[a] De Nicola: Dipartimento di Ingegneria dell'Impresa. Università di Roma "Tor Vergata", Via del Politecnico 1, 00133 Rome, Italy
[b] Gitto: Dipartimento di Ingegneria dell'Impresa. Università di Roma "Tor Vergata", Via del Politecnico 1, 00133 Rome, Italy
[c] Mancuso: Dipartimento di Ingegneria dell'Impresa. Università di Roma "Tor Vergata", Via del Politecnico 1, 00133 Rome, Italy

ARTICLE INFO

Keywords:
Bootstrap
Data envelopment analysis (DEA)
Classification and regression trees
Environmental variables
Health policy
Efficiency
Patient mobility

ABSTRACT

One of the main problems in efficiency analysis is to determinate the environmental variables that have an impact on the production process. This paper shows that applying bootstrap to data envelopment analysis (DEA) before performing classification and regression trees (CART) increase the quality of the results. In particular, employing data on the Italian Health System, the paper highlights that bias corrected DEA allows to individuate variables affecting health efficiency which would remain undiscovered when the traditional DEA model is applied.

1. Introduction

Many models have been developed to find an optimal solution to the problem to improve healthcare efficiency. In this paper, the concept of efficiency, measured by data envelopment analysis (DEA), is implemented together to classification and regression trees (CART) analysis to provide a set of rules that permit to identify on which environmental variables the governments should operate to improve healthcare efficiency. DEA is a well know non-parametric method developed by Charnes, Cooper, and Rhodes (1978) that identifies a production frontier and determines the efficiency scores of a set of decision making units (DMU), with the common set of inputs and outputs (Heidari & Mohammadi, 2012; Lin, Lee, & Chiu, 2009). In the other hand, one of the significant limits on applying this non-parametric technology is that the efficiency scores are an estimate of the true (and unknown) production frontier, conditional on observed data resulting from an underlying Data Generating Process (DGP) (Simar & Wilson, 1998, 2000). As a consequence, DEA efficiencies are biased by construction and are sensitive to the sampling variations of the obtained frontier. In order to overcome this problem, Simar and Wilson (1998) proposed a bootstrap procedure to approximate the sampling distribution of the efficiency scores and to make inference. See Halkos and Tzeremes (2012), Curi, Gitto and Mancuso (2011) and Gitto and Mancuso (2012) for recent applications of bootstrap-DEA methodology.

The CART methodology (Breiman, Friedman, Olshen, & Stone, 1984) which allows to identify some rules with the aim to classify a sample into two or more groups, has been applied in different fields (D'uva & De Siano, 2007; Li, Sun, & Wu, 2010; Sohn & Tae, 2004). Nowadays, to the best of our knowledge, it is never applied to support policy intervention in the health system.

In this paper, bootstrapped DEA and CART analysis are implemented in order to define policy intervention aimed to improve health efficiency. Moreover, this study discusses the importance to use DEA in an inferential setting by employing the bootstrap technique.

1.1. Research objectives

The main objectives of this study are:

1. To demonstrate the applicability of the CART methodology in the health sector.
2. To stress the importance of the bootstrap in DEA analysis.

2. The proposed methodology

The methodology used in the paper is illustrated in Fig. 1. It is composed by two different stages; the first deals with DEA technique while the second concerns the use of CART technique.

2.1. Data envelopment analysis

DEA is an efficiency evaluation approach entirely based on the observed data. The main concept is that the efficiency of a specific DMU is determined by its capability to obtain desirable outputs

* Corresponding author. Tel.: +39 0672597793; fax: +39 0672597305.
E-mail addresses: arianna.de.nicola@uniroma2.it (A. De Nicola), simone.gitto@uniroma2.it (S. Gitto), paolo.mancuso@uniroma2.it (P. Mancuso).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.02.074

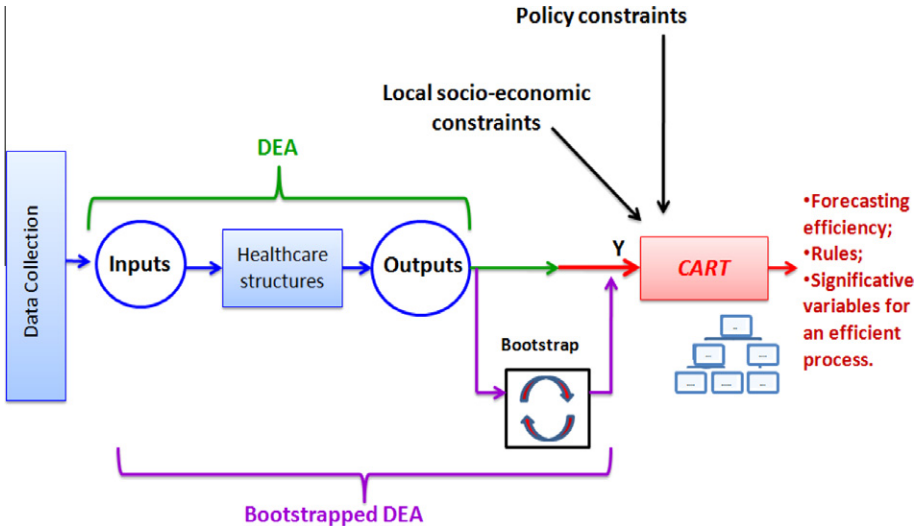

Fig. 1. The methodology.

from a set of inputs. So, this methodology constructs an efficient production frontier based on the best practice, applying a linear programming technique to the established sample. In order to facilitate the interpretation of the results in the next sections, it is useful to recall that in the output orientated DEA model, under the hypothesis of variable return to scale (VRS), an efficiency score $\hat{D}_{it}$ is calculated for each DMU i ($i = 1,2,\ldots,n$) at time t (t=1,2,…,T), by solving the following linear program:

$$\begin{aligned} &\hat{\theta}_{it} = [\hat{D}_{it}]^{-1} = \max_{\theta\lambda} \theta \\ &s.t. x_{it} \geqslant X_t \lambda \\ &\theta y_{it} \leqslant Y_t \lambda \quad i = 1,2\ldots,n; \quad t = 1,2\ldots,T \\ &1'\lambda = 1 \\ &\lambda \geqslant 0 \end{aligned} \tag{1}$$

where $\hat{\theta}_{it}$ and $\hat{D}_{it}$ are the Farrell (1957) and Shepard (1970) distance functions, n is the number of DMUs and T is the number of time periods; $\mathbf{Y}_t$ is a *sxn* matrix of s outputs, $\mathbf{X}_t$ is a *rxn* matrix of r inputs, λ represent a *nx1* vector of weights which allows to obtain a convex combination between inputs and outputs and **1'** is a vector of ones.

Now, $\hat{\theta}_{it}$ is an inefficiency measure and always assumes values equal to or greater than one. Consequently, $\hat{D}_{it}$ is an efficiency measure and it assumes values between zero and one. DMU with an efficiency score equal to one are located on the frontier and as consequence their outputs cannot be further expanded without a corresponding increase in inputs.

In the first stage of this analysis, we assume an output-orientated model with variable return to scale to maximize the outputs that could be produced given the inputs (Ancarani, Di Mauro, & Giammanco, 2008; Barbetta, Turati, & Zago, 2007; Ferrier, Rosko, & Valdmanis, 2006).

The DEA approach offers many strengths: minimum assumptions about the structure of production, flexibility and direct relationship to the economic theory (Coelli et al.,1998).

2.2. The bootstrap DEA

Nevertheless, as discussed by Simar and Wilson (2000), DEA estimator is biased toward unity. In fact, relation (1) does not allow us to determine whether the efficiency values are real, or merely an artifact of the fact that we do not know the true production frontier and must estimate them from a finite sample (Simar & Wilson, 2000). In a context of two-stages procedure as proposed in this paper, the use of biased scores can lead to misleading results as discussed by Simar and Wilson (2007). Consequently bootstrapping techniques, based on the idea that the DGP can be estimated by using the given sample to generate a set of bootstrap samples from which parameters of interest can be calculated, must be used to obtain unbiased results. In the research results, we show what happens in a case study, when the bias is not taken into account.

Following Simar and Wilson (1998), we employ a consistent bootstrap estimation procedure to obtain the sampling distribution of the efficiency scores, and so to correct for the bias. The idea underlying the bootstrap is to approximate the sampling distributions of $\hat{\theta}$, by simulating their DGP. In other terms, given the estimates $\hat{\theta}_{it}$ of the unknown true values of θ_{it} we generate through the DGP process a series of bootstrap estimates $\hat{\theta}^*_{it}$. Thus, for the generic unit i, compute the bias term:

$$BIAS(\hat{\theta}_i) = B^{-1} \sum_{b=1}^{N} \hat{\theta}^*_{i,b} - \hat{\theta}_i, \quad \forall\, i = 1,\ldots,n \tag{2}$$

where $\hat{\theta}^*_{i,b}$ is the bootstrapped technical efficiency and B is the number of bootstrap replications. The bias-corrected estimator of $\hat{\theta}_i$ is:

$$\hat{\theta}^c_i = \hat{\theta}_i - BIAS(\hat{\theta}_i) = 2\hat{\theta}_i - B^{-1} \sum_{b=1}^{B} \hat{\theta}^*_{i,b} \tag{3}$$

The quality of the bootstrap depends on both the number of replications and the sample size (Simar & Wilson, 2000). If the bootstrap is consistent, then:

$$(\hat{\theta}_{it} - \theta_{it}) \,|\hat{s} \quad \sim^{approx} \quad (\hat{\theta}^*_{it} - \hat{\theta}_{it}) \,|S^* \quad i = 1,2,\ldots,n \quad t = 1,2,\ldots,T \tag{4}$$

where, $\hat{s}$ and S^* denotes the observed and the bootstrap sample.

In the case study, the results of the model are obtained from 2000 iterations.

2.3. Classification and regression tree (CART)

In the second step, we use a regression type CART where the explanatory variables represent the characteristics of population and health services provided. CART, is a non-parametric statistical procedure for predicting a dependent variable using some explanatory variables (predictors). In particular, the major goal of this methodology is to uncover the predictive structure of the health efficiency in the Italian provinces, creating an accurate dataset. So CART algorithm permits, by binary recursive partitioning, to find through all value of predictors, those minimizes the weighted variance (Razi & Athappilly, 2005). The final tree consists of a root node that includes all the observations, some parent nodes which can be

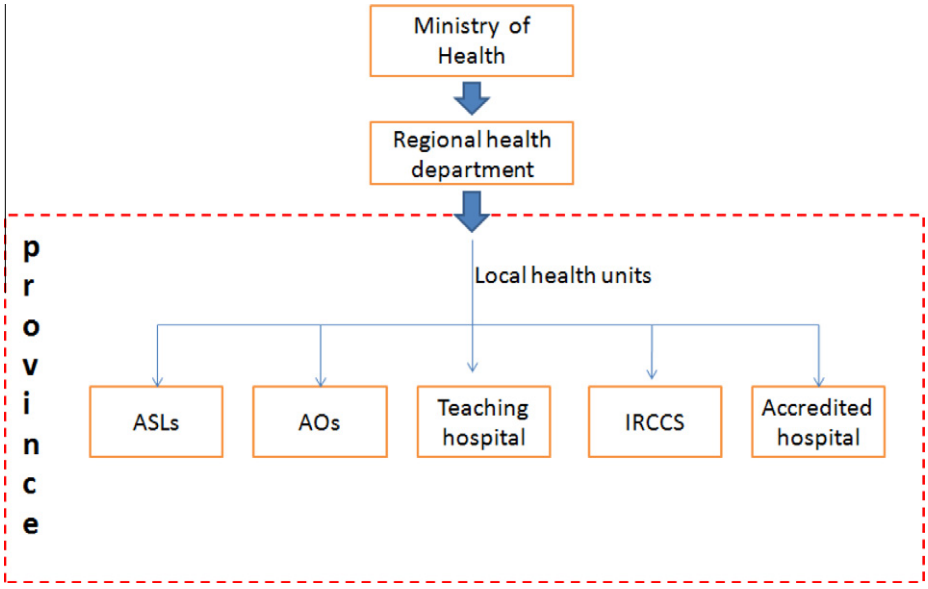

Fig. 2. The Italian national healthcare system.

Table 1
Descriptive statistics for DEA inputs and output.

Variables	Description	Mean	Std. Dev.	Median
Physicians	Number of salaried physicians and dentists	1697	2069.19	1098
Beds	Number of beds	1807	2142.65	1128
Nurses	Number of paid nurses	3993	4112.30	2788
Ordinary discharge	Number of ordinary discharged adjusted for CMI	68940	85360.11	41510

still splitted and, at the end of the tree, some terminal nodes (leaves). Each leave contains set of observations and is characterized by an average value that represents the predicted value of the dependent variable. So, the final tree is characterized only by the explanatory variables that are predictive for the dependent variable. However, the same explanatory variable can be used several times in different levels of the tree. After building the largest possible tree, the pruning is recommended, to increase the predictive accuracy of the CART. Tree optimization is implemented using the cross-validation. The procedure has been implemented using the statistical software R with the libraries FEAR and RPART.

3. Case study

3.1. Healthcare organization in Italy

The Italian National healthcare system (SSN) is founded by tax revenue and provides universal coverage to all Italian citizen and foreign legal residents. At the beginning, the SSN was strongly centralized but recently, the Central Government has radically changed its structure to a more decentralized one. The Italian SSN is made up of three decision makers: (1) the central government with the Ministry of Health, (2) the 21 regional governments and (3) the local health units that comprise a number of organization as local health enterprises (*Aziende Sanitarie Locali*, ASL), public hospital enterprises (*Aziende Ospedaliere*, AO), national institutes for scientific research and private accredited providers (see Donatini et al., 2001; Lo Scalzo et al., 2009 for a more extensive discussion). Fig. 2 shows in a synthetic way the SSN structure. The local units can provide different health services to the local population that differ mostly for kind and quality. This aspect might involve a non homogeneous comparison among the DMUs, so we merge all the local health units at the provincial level. The province is an intermediate bureaucratic institution between municipality and region. It is employed by the regional government to plan a comprehensive set of health services covering the area as a whole.

3.2. Data collection

To evaluate the possible relationships among healthcare efficiency and environmental variables, this paper analyzes as case study 98 Italian provinces in 2006, for which the data were collected from the Italian Ministry of Health[1] and from National Institute for Statistics (Istat).

In order to estimate healthcare efficiency of our sample by DEA, three inputs (physicians, nurses and beds) and one output (ordinary discharge) were selected. All inputs and output are measured in terms of physical quantities, since no reliable price data are available.

Descriptive statistics for the variables used in the DEA analysis are reported in Table 1.

In this paper we select six explanatory variables to be used in CART analysis: patient inflow (p_inf), patient outflow (p_out), deprivation index (DEP), the percentage of ASL hospital beds (ALS_beds), the percentage of caesarean (CAE) and the beds utilization rate (UT_BEDS).

The descriptive statistics for these variables are resumed in Table 2. The variables *p_inf* and *p_out* measure patient movements among provinces for healthcare reasons. Analyzing the patient mobility between health organizations allow us to understand if policy decision with the aim to regulate this aspect might represent a real solution to improve health efficiency.

[1] www.salute.gov.it.

Table 2
Summary of the explanatory variables.

Variables	Description	Max	Min	Mean	Std. Dev
Patient inflow (p_inf)	Patients who enter for health reason from another province (rate)	49.27	6.65	19.25	7.98
Patient outflow (p_out)	Patient that exit to the province for health reason (rate)	44.20	4.60	22.99	9.65
Deprivation Index (DEP)	Measure the socio-economic disadvantage to live in a particular area	8.778	−2.901	0.024	1.61
percentage of caesarean (CAE)	The percentage of caesarian births in each province (%)	65.06	14.34	35.99	28.45
ASL hospital beds (ALS_beds)	The percentage of hospital beds directly managed by the region in each provinces (%)	1	0	0.5016	0.31
beds utilization rate (UT_BEDS)	The beds occupancy rate in each province (rate)	89.20	62.95	79.20	4.98

Considering that new reforms seem to change again the regional financial autonomy in the SSN, taking into account also socio-economic disparities may be important to guarantee the equity of health services provided. So, in this financial prospective, a better allocation of the healthcare resources should consider also the social disadvantages related to the population characteristics that might characterized the country. This is possible considering the deprivation index. The main idea is that, according to the equity consideration, more deprived areas need a larger amount of resources. So the deprivation index was constructed considering different aspect of disadvantages that might affect the population: education, employment, housing and domestic condition. In particular, we construct the deprivation index following Caranci, Spadea, and Costa (2009) and using Istat data from census of 2001.[2]

The last wave of reforms has given a greater independence to the regions about health policy. It has made possible a great deal of variation in how each region performs its role of 'third party payer/ purchaser' in the healthcare system. In fact, recent investigations of the Italian national health system highlight that there are important differences among the regions about their governance capability and this seems to influence directly cost and quality of services provided (Mapelli, 2007). In addition, Francese, Piacenza, Romanelli, and Turati (2010) discuss that policy outcomes may be influenced also by the way in which the levels of governments interact each other, as suggested by the modern theory of federalism. So, to analyze how each region applies its role of 'third party payer/ purchaser', we use *ASL_beds variable* (Lo Scalzo et al. 2009; Mapelli, 2007).

The variable *CAE* is considered as an indicator of (in) appropriateness of the health performance. In fact, it is a surgical treatment characterized by higher cost, respect to the classical vaginal delivery, so in absence of any therapeutic reason, this treatment is usually considered an inappropriate way of delivery (Francese et al. 2010). For the policy makers, the reduction of the available resource and the need to reduce health expenditures leads to limiting the inappropriateness. As a consequence, to reduce costs may represent one of the most important methods to increase health efficiency and to improve resource allocation.

The variable *UT_BEDS* is considered in order to the effective utilization of the available resources in the provinces. In fact, recent policy decisions have forced, in the short term, the hospital organizations to keep the amount of resources consumed constant and, at the same time, to increase health services provided. So monitoring also this aspect may allow a better allocation of the resources.

4. Research results

In order to verify if applying a bootstrapped DEA implies significant differences in the selection of the environmental variables on which operate to improve health efficiency, we propose two models that differ only for the bias correction to the DEA technique. Then, using a CART algorithm we obtain two tree-shaped structures.

[2] http://dawinci.istat.it/MD/.

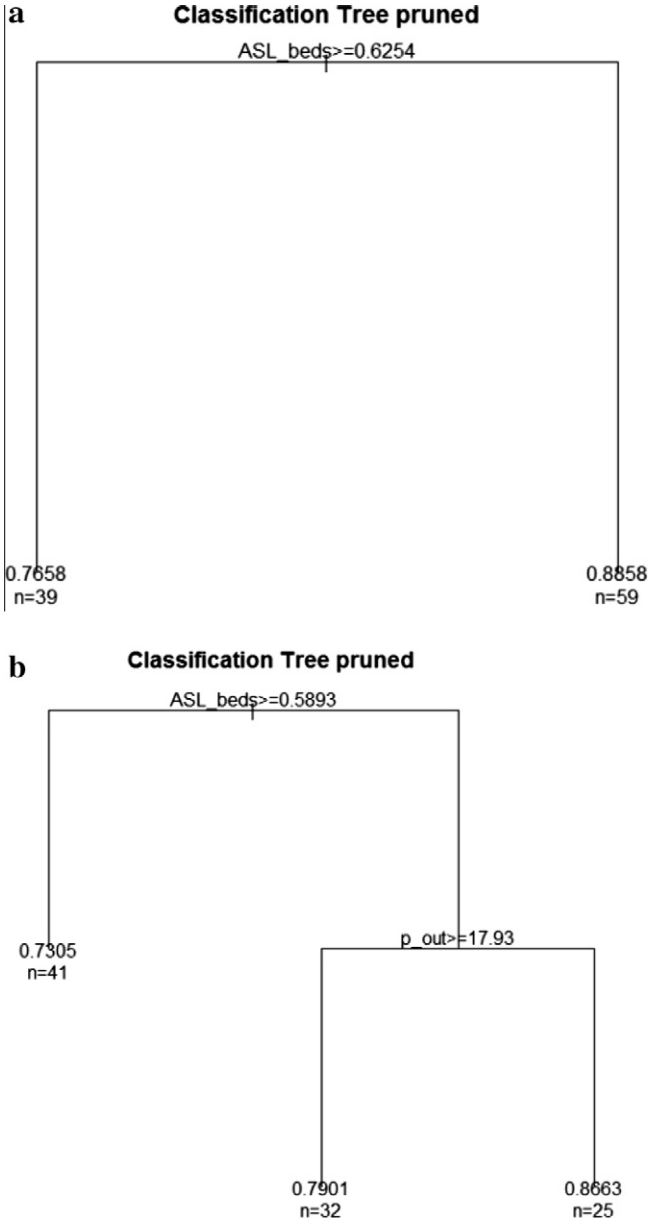

Fig. 3. a. Pruned tree of DEA. b. Pruned tree of bias corrected DEA.

Fig. 3 shows the results of the pruned trees where the specific path, starting from the root node to the terminal node, characterized each specific group.

The results confirm that the *percentage of ASL hospital beds* directly managed by the region is the most important variables in order to classify the efficiency of the structure. These results imply that the way in which the region performs its role of "third player" is strictly related to the performance of the health system. In fact, the analysis highlights that a direct control of the regional government on hospital beds, seems do not influence positively the health

performance. The main difference between the two models is that, when we correct for the bias using the bootstrap, also the patient outflow becomes an important variable on which operate to classify the efficiency of the structure. In fact, regulating the patient outflow under specific threshold value seems to represent a rule for limiting inefficiency in health system. So, it is observed that the implementation of the bootstrap-DEA procedure in a CART analysis involves more accurate results, allowing in this way a better planning of the reforms aimed to improve health efficiency.

5. Conclusion and future work

DEA has had wide applications in measuring the relative efficiency scores of a sample of DMUs. But, especially in health sector, the performance is often affected by environmental variables on which government and managers hardly operate. This paper has employed a bootstrap-DEA methodology with CART analysis as the tool for uncover the predictive structure of health efficiency. The paper pointed out that, the percentage of hospital beds directly managed by the regional government, used as proxy of the way in which the region performs its role of "third party player/purchaser" in the health direction, results to be the environmental variable with the most influential role in determining health efficiency. This imply that the regional governments, promoting a significant form of competition inside their healthcare system, can increase their health efficiency. In addition, the analysis demonstrate that future policy intervention, aimed to increase the health efficiency, should also consider patient mobility. Finally, under a methodological perspective, the paper underlines that the use of bias corrected efficiency scores, improve the quality of the CART analysis.

Appendix A. Appendix 1

In particular, following Caranci et al. (2009) we construct the deprivation index (DEP), considering the following variables:

- $X1$ = [Population with primary education, literate and illiterate/population at least six year] $*$ 100.
- $X2$ = [(man power − unemployed or job seekers)/man power] $*$ 100.
- $X3$ = [houses occupied by resident in rent/houses occupied by resident] $*$ 100.
- $X4$ = [Population/area (m^2) houses occupied by resident] $*$ 100.
- $X5$ = [lone parent with childhood/Total family] $*$ 100.

The index is an unweighted combination of five standardized variables; let

$Z_i = \frac{x_i - m_{xi}}{s_{xi}}$ being m_i = indices means $i = 1 \ldots 5$, and s_i = indices standard deviations; the deprivation index is given by: ID = $\sum_{i=1}^{5} z_i$. These measures are calculated for all the Italian provinces.

References

Ancarani, A., Di Mauro, C., & Giammanco, M. D. (2008). The impact of managerial and organizational aspects on hospital wards' efficiency: Evidence from a case study. *European Journal of Operational Research, 194*, 280–293.

Barbetta, G. P., Turati, G., & Zago, A. M. (2007). Behavioral differences between public and private not-for-profit hospitals in the Italian National health service. *Health Economics, 16*, 75–96.

Breiman, L., Friedman, J., Olshen, R., & Stone, J. (1984). *Classification and regression trees*. Belmont, California: Wadsworth.

Caranci, N., Spadea, T., & Costa, G. (2009). Deprivazione e mortalità. *Rapporto Osservasalute*, 41–47.

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

Coelli, T., Rao, D. S. P., & Battese, G. E. (1998). *An introduction to efficiency and productivity analysis*. Boston: Kluwer Academic Publishers. Inc.

Curi, C., Gitto, S., & Mancuso, P. (2011). New evidence on the efficiency of Italian airports: A bootstrapped DEA analysis. *Socio-Economic Planning Science, 45*, 84–93.

D'uva, M., & De Siano, R. (2007). Human capital and "club convergence" in Italian regions. *Economics bulletin, 18*, 1–7.

Donatini, A., Rico, A., D'Ambrosio, M. G, Lo Scalzo, A., Orzella, L., Cicchetti, A., et al. (2001) *Health Care Systems in Transition: Italy*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Farrell, M. J. (1957). The measurement of productive efficiency. *Journal of Royal Statistical Society A, 120*, 253–281.

Ferrier, G. D., Rosko, M. D., & Valdmanis, V. (2006). Analysis of uncompensated hospital care using a DEA model of output congestion. *Health Care Management Science, 9*, 181–188.

Francese, M., Piacenza, M., Romanelli, M., & Turati, G. (2010). Understanding Inappropriateness in Health Treatments: The Case of Caesarean Deliveries across Italian Regions. In *Working paper, University of Torino, Department of Economics and Public Finance "G. Prato"*.

Gitto, S., & Mancuso, P. (2012). Bootstrapping the Malmquist indexes for Italian airports. *International Journal of Production Economics, 135*, 403–411.

Halkos, G. E., & Tzeremes, N. G. (2012). Industry performance evaluation with the use of financial ratios: An application of bootstrapped DEA. *Expert System with Applications, 39*, 5872–5880.

Heidari, M. D., & Mohammadi, M. O. A. (2012). Measuring productive efficiency of horticultural greenhouses in Iran: A data envelopment analysis approach. *Expert Systems with Applications, 39*, 1040–1045.

Li, H., Sun, J., & Wu, J. (2010). Predicting business failure using classification and regression tree: An empirical comparison with popular classical statistical methods and top classification mining methods. *Expert Systems with Applications, 37*, 5895–5904.

Lin, T. T., Lee, C.-C., & Chiu, T.-T. (2009). Application of DEA in analyzing a bank's operating performance. *Expert Systems with Applications, 36*, 8883–8891.

Lo Scalzo, A., Donatini, A., Orzella, L., Cicchetti, A., Profili, S., & Maresso, A. (2009). *Health Care Systems in Transition: Italy*. Copenhagen: WHO Regional Office for Europe on behalf of the European Observatory on Health Systems and Policies.

Mapelli, V. (Ed). (2007). I Sistemi di Governance dei Servizi Sanitari Regionali [The System of Governance of Regional Health Services]. Roma: Formez, Quaderni, n. 57, Available at <http://sanita.formez.it/>.

Razi, M. A., & Athappilly, K. (2005). A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications, 29*, 65–74.

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

Simar, L., & Wilson, P. W. (1998). Sensitivity analysis of efficiency scores: How to bootstrap in non-parametric frontier models. *Management Science, 44*, 49–61.

Simar, L., & Wilson, P. W. (2000). Statistical inference in non-parametric frontier models: The state of art. *Journal of Productivity Analysis, 13*, 49–78.

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

Sohn, S. Y., & Tae, M. H. (2004). Decision tree based on data envelopment analysis for effective technology commercialization. *Expert Systems with Applications, 26*, 279–284.