Available online at www.sciencedirect.com
ScienceDirect
European Journal of Operational Research 187 (2008) 629–635
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Interfaces with Other Disciplines

Outlier detection in two-stage semiparametric DEA models

Andrew L. Johnson [a,*], Leon F. McGinnis [b]

[a] Department of Industrial and Systems Engineering, Texas A&M University, College Station, TX 77840, USA
[b] Department of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

Received 9 November 2005; accepted 26 March 2007
Available online 21 April 2007

Abstract

In the use of peer group data to assess individual, typical or best practice performance, the effective detection of outliers is critical for achieving useful results, particularly for two-stage analyses. In the DEA-related literature, prior work on this issue has focused on the efficient frontier as a basis for detecting outliers. An iterative approach for dealing with the potential for one outlier to mask the presence of another has been proposed but not demonstrated. This paper proposes using both the efficient frontier and the inefficient frontier to identify outliers and thereby improve the accuracy of second stage results in two-stage nonparametric analysis. The iterative outlier detection approach is implemented in a leave-one-out method using both the efficient frontier and the inefficient frontier and demonstrated in a two-stage semi-parametric bootstrapping analysis of a classic data set. The results show that the conclusions drawn can be different when outlier identification includes consideration of the inefficient frontier.

Keywords: Data envelopment analysis; Productivity; Inefficient frontier; Outlier detection

1. Introduction

Productivity and efficiency have long been research areas for both economists and engineers. Productivity is the ratio of outputs produced to inputs consumed and efficiency is the ratio of a given system's productivity compared to the best possible productivity, Lovell (1993). Many models have been proposed for determining the best possible productivity. A main concern in constructing these models or evaluating them is whether the productivity identified is truly achievable for the system under consideration. This has lead researchers to investigate and quantify the effects on efficiency by the environment and other variables that cannot be controlled by system management. One of the most common types of models for this purpose has come to be known as two-stage semi-parametric models, first suggested by Timmer (1971).

In the first stage a deterministic frontier model is constructed. When the assumptions of convexity and free disposability are made, this calculation is referred to as data envelopment analysis (DEA), Charnes et al. (1978). Other deterministic frontier techniques also may be used, such as the free disposal hull (FDH) first rigorously analyzed by Deprins et al. (1984). In the second stage, the efficiency estimates calculated in the first stage are regressed against a variety of environmental variables. The

* Corresponding author. Tel.: +1 9798459025.
E-mail address: ajohnson@tamu.edu (A.L. Johnson).

0377-2217/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2007.03.041

first implementations of the two-stage semi-parametric models were by Ray (1988) and Ray (1991). However, these methods have recently been criticized by Simar and Wilson (2007) for their lack of a coherent data-generating process and mishandling of the complicated unknown serial correlation among the estimated efficiencies.

Wilson (1995) and others note that in the first stage, the deterministic nature of the frontier implies that errors in measurement for those observations supporting the frontier could cause severe distortions in the measures of efficiency for the entire population. Wilson then suggests a method to remedy this problem by calculating the leave-one-out efficiency, sometimes called super efficiency, and identifying outliers based on the leave-one-out efficiency estimate. The leave-one-out efficiency estimate has been suggested by other researchers, including Banker et al. (1989), Anderson and Petersen (1993), and Lovell et al. (1993). The Banker and Das paper refers to its use for outlier measurement whereas the latter two papers use the method for tie breaking among the observations that appear to be efficient. Wilson relates the problem of identifying observations with measurement error to the problem of outlier detection in the classical linear regression models. However, outliers in linear regression models can be found both above and below the regression line, whereas, Wilson's method only identifies a subset of outliers related to being "too good" or to continue the regression analogy, outliers found above the regression line.

While outliers are an intuitive concept, a rigorous definition is hard to state. Assuming data have been generated by drawing from a distribution, an observation categorized as an outlier actually may simply represent a low probability draw (i.e., a draw from a tail of the distribution). While such a draw may appear to be an outlier, as Cook and Weisberg (1982) point out, this type of observation may lead to the recognition of important phenomena that might otherwise go unnoticed. With this in mind the rather loose definition of outlier provided by Gunst and Mason (1980), "as observations that do not fit in with the pattern of the remaining data points and are not at all typical of the rest of the data", seems appropriate. In deterministic frontier models, outliers that support the frontier can be thought of as observations that are "too good" and thus are particularly dangerous as noted above by Wilson. The observation motivating this work is: *when the two-stage semi-parametric model is used, outliers that represent particularly bad performance might distort the second stage results.*

There has not been much research in the area of identifying outliers relative to a nonparametric deterministic frontier. There appear to be no published literature discussing how to identify outliers which distinguish themselves by exhibiting particularly poor performance. The available research (Wilson, 1995; Simar, 2003) focuses only on identifying outliers which impact the efficient frontier. Many studies have been performed to measure sensitivity or robustness of DEA results and while this is closely related to many techniques for identifying outliers, the concept is fundamentally different. There has been limited attention paid to inefficient frontiers. Paradi et al. (2004) suggest a worst practice detection method by applying traditional DEA models when only detrimental (bad) outputs are selected. In their approach, a new mathematical formulation is not needed; poor performers are simply identified by high levels of bad outputs. Liu and Hsu (2004) also have suggested a similar mathematical formulation for identifying an inefficient frontier; however, the paper provides no motivation for developing an inefficient frontier.

The present paper describes an inefficient frontier and how this concept can be used to identify outliers that distinguish themselves by having particularly poor performance. Clearly, observations identified as potential outliers should be further examined to determine if an error has taken place, possibly in data entry or in identifying these units as members of the peer group for this analysis. This paper also describes the implementation of the iterative outlier identification process, discussed in Wilson (1995) although apparently not demonstrated in the literature. Section 2 will review the two-stage semi-parametric models using data envelopment analysis (DEA) in the first stage and bootstrapping methods in the second stage. Section 3 will address methods for constructing an inefficient frontier and describe outlier detection methods applied to the inefficient frontier. An example using the classic Banker and Morey (1986) data set will be shown in Section 4. The impact on second stage results of not identifying and processing inefficient outliers will be demonstrated. Finally, conclusions will be presented.

2. Two-stage semi-parametric bootstrapping method

The two-stage semi-parametric model approach consists of estimating efficiencies in the first-stage

and regressing these efficiency estimates against a set of environmental variables in the second-stage. Many models are available for estimating efficiency; but we will focus on the DEA model. The DEA production set can be described by

$$\widehat{P} = \{(x,y)|y \leqslant Y\lambda,\ x \geqslant X\lambda,\ i^{\mathrm{T}}\lambda = 1,\ \lambda \in R_{+}^{n}\}, \tag{2.1}$$

where $\widehat{P}$ is an estimate based on the observed pairs (x_i,y_i) of the "true" production set P, $x \in R_{+}^{p}$ denotes a $(1\times p)$ vector of inputs, $y \in R_{+}^{q}$ denotes a $(1\times q)$ vector of outputs, n is the number of observations, $Y = [y_1 \ldots y_n]$, $X = [x_1 \ldots x_n]$, i denotes an $(n\times 1)$ vector of ones, and λ is an $(n\times 1)$ vector of intensity variables. The production set can be completely described by either the input requirements set or the output requirement set. The input set can be stated as

$$L(y) = \{x \in R_{+}^{p} | x \text{ can produce } y\}. \tag{2.2}$$

To simplify exposition, we will focus on the input space, however, the concepts described for the input space transfer easily to the output space. For further description of the relationship between the two spaces see either Lovell (1994) or Charnes et al. (1993). The linear program for calculating the efficiency estimates in the input requirement space is

$$\begin{aligned} \min_{\hat{\theta}_i,\lambda} \quad & (\hat{\theta}_i), \\ \text{s.t.} \quad & -y_i + Y\lambda \geqslant 0, \\ & \hat{\theta}_i x_i - X\lambda \geqslant 0, \\ & \sum_{j=1}^{N} \lambda_j = 1, \\ & \lambda_j \geqslant 0. \end{aligned} \tag{2.3}$$

This linear program is solved once for each observation, $i = 1,\ldots,n$ to compute efficiency estimates for that observation.

Let $z \in R_{+}^{r}$ denote a $(1\times r)$ vector of environmental variables. In a two stage analysis, a function, typically $\psi(z_i,\beta) = z_i\beta$ is specified and an associated regression model is

$$\hat{\theta}_i = z_i\beta + \varepsilon_i, \tag{2.4}$$

where ε_i is an error term, normally distributed and truncated so that the values of $\hat{\theta}_i$ do not exceed 1. $\hat{\theta}_i$ is the efficiency estimated from the first-stage for observation i. The sign on the resulting coefficients, β, indicate the direction of the influence and hypothesis testing can assess the significance. Until recently this had been standard practice and advocated by several researchers, e.g., Coelli et al. (1998), McCarty and Yaisawarng (1993), and Ray (1991).

In 2007 Simar and Wilson introduced a bootstrapping technique to improve the inference in the second-stage regression. They cited a two fold need for a new technique as: (1) the original two-stage method lacks a coherent data-generating process; and (2) it mishandles the complicated unknown serial correlation among the estimated efficiencies and the correlation between the ε_i and the z_i.

The Simar and Wilson bootstrapping technique uses Shephard's input distance function which is inversely related to an input efficiency estimate. Shephard's input distance function is

$$D_i(x_i,y_i) = (\hat{\theta}_i)^{-1} = \max\left\{\hat{\delta}_i \middle| \frac{x_i}{\hat{\delta}_i} \in L(y)\right\}. \tag{2.5}$$

The value of $\hat{\delta}_i$ is a normalized measure of the distance from a point (x_i,y_i) to the frontier, holding output levels and the direction of the input vector fixed.

The DEA method of using a set of observation to approximate the efficient frontier, biases efficiency estimates upwards. This bias in the efficiency estimates is another reason why the bootstrapping technique is necessary. In the remainder of this paper, algorithm #2 from Simar and Wilson (2007) is used to perform the second stage regression. The bootstrapping method is an improvement over the original deterministic frontier two-stage model because the associated confidence intervals allow one to quantify the uncertainty related to efficiency estimates. However, before the two-stage model can be implemented, unexplainable outliers should be identified and removed from the data. The previous outlier detection methods were only concerned with the overly efficient outliers because they were developed with the traditional deterministic frontier model in mind and did not consider the second stage regression. Thus the outliers that were overly inefficient would have minimal impact on their results. However, when the two-stage model is considered the effect of overly inefficient outliers may cause misleading results in the second stage, as will be shown in Section 4.

3. The inefficient frontier, outliers, and a detection methodology

The outlier detection methodology for non-parametric efficiency evaluation described here is

distinguished from previous methodologies by searching for both efficient and inefficient outliers. In order to identify inefficient outliers a measure to quantify a given observation's deviation from the remainder of the data set needs to be defined. This is done using the concept of the inefficient frontier, introduced below. The method that will be used to identify outliers is the leave-one-out method described by Wilson (1995), applied to both the efficient and inefficient frontiers.

For the purposes of this paper, we assume that clustering or other techniques, if appropriate, already have been applied to obtain data in a comparable group.

3.1. The inefficient frontier

Just as an efficient frontier can be calculated from observations taken from the production set P, an inefficient frontier also can be calculated. The efficient frontier represents the maximum output given an input level and without improvements in technology it is not possible to achieve greater production levels. By analogy, the deterministic inefficient frontier can be defined, from the output perspective as, a convex hull defined by the minimum output level given an input level, for which it would not be likely to produce output levels less than the frontier value. Convex combinations of the most inefficient observed units estimate the inefficient frontier, in a manner analogous to the estimation of the efficient frontier. Similarly, from the input perspective the inefficient frontier is a convex hull defined by the maximum input level given an output level, for which it would not be likely to use input levels greater than the frontier value. An observation may lie outside of the inefficient frontier if there is error in the measurement or entry of the data, if the observation is a chance instance of a low probability situation, or if the observation does not truly belong to the group under evaluation. For any of these reasons a data point should be removed from the analysis. When outlier detection techniques are applied to the inefficient frontier, the envelopment concept is relaxed in order to quantify the degree to which each unit on the inefficient frontier (call these units completely inefficient units) is an outlier. These units represent the worst possible performance within the observed production possibility set.

When the inefficient frontier is included the production possibility set is defined as:

$$\widehat{P}_p = \left\{ \begin{array}{l} (x,y) | y \leqslant Y\lambda, x \geqslant X\lambda, i'\lambda = 1, \lambda \in R_+^n \\ \text{and } y \geqslant Y\mu, x \leqslant X\mu, i'\mu = 1, \mu \in R_+^n \end{array} \right\}. \tag{3.1}$$

The input set can be stated as

$$L_p(y) = \{x \in R_+^p | x \text{ can produce } y\}, \tag{3.2}$$

given the new definition of production possibility set and the output set as

$$K_p(x) = \{y \in R_+^q | y \text{ can be produced by } x\}, \tag{3.3}$$

given the new definition of production possibility set. For this new definition, a Shephard's input inefficient distance function can be defined:

$$D_{i\mathrm{II}}(x_i, y_i) = \min\{\psi_{i\mathrm{I}} | x_i/\psi_{i\mathrm{I}} \in L_p(y)\}, \tag{3.4}$$

where the subscripts on D indicate unit, input and inefficiency, respectively. $D_{i\mathrm{II}} \leqslant 1$ with 1 characterizing the inefficient frontier. Similarly, a Shephard's output inefficient distance function can be defined as

$$D_{i\mathrm{OI}}(x_i, y_i) = \max\{\psi_{i\mathrm{O}} | y_i/\psi_{i\mathrm{O}} \in K_p(x)\} \tag{3.5}$$

$D_{i\mathrm{OI}} \geqslant 1$ with 1 characterizing the inefficient frontier.

The shape of the one-input, one-output inefficient frontier shown in Fig. 1 is an approximation of the true inefficient frontier, constructed from the observed data.

The inefficient frontier with respect to the subset $L_p(y)$ can be denoted as $\partial X_{\mathrm{in}}(y)$ and found by

$$\partial X_{\mathrm{in}}(y) = \{x | x \in L_p(y), \phi_{\mathrm{I}} x \notin L_p(y) \forall 1 < \phi_{\mathrm{I}}\}. \tag{3.6}$$

Then the inefficiency estimate calculated from the input perspective can be found by solving the following linear program

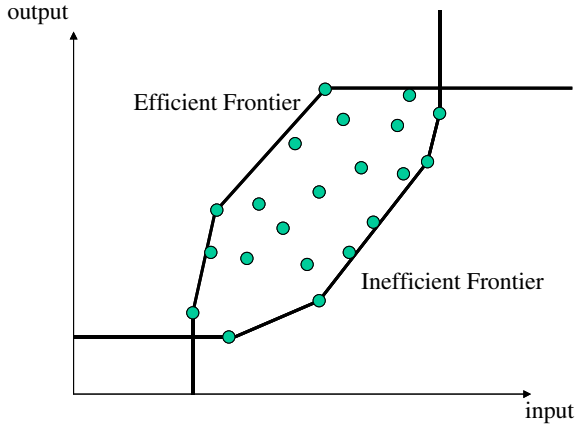

Fig. 1. The inefficient and efficient frontiers for one input, one output.

$$\begin{aligned} \max_{\phi_{i\mathrm{I}},\mu\lambda} \quad & (\phi_{i\mathrm{I}}), \\ \text{s.t.} \quad & -y_i + Y\mu \leqslant 0, \\ & \phi_{i\mathrm{I}} x_i - X\mu \leqslant 0, \\ & \sum_{j=1}^{N} \mu_j = 1, \\ & \mu_j \geqslant 0. \end{aligned} \tag{3.7}$$

For completeness, the inefficient frontier with respect to the subset $K_p\ (x)$ can be denoted as $\partial Y_{\mathrm{in}}(x)$ and found by

$$\partial Y_{\mathrm{in}}(x) = \{y | y \in K_p(x),\ \phi_{\mathrm{O}} y \notin K_p(x) \forall 0 < \phi_{\mathrm{O}} < 1\}. \tag{3.8}$$

We also show the linear program for calculating the inefficiency estimate from the output perspective

$$\begin{aligned} \min_{\phi_{i\mathrm{O}},\mu} \quad & (\phi_{i\mathrm{O}}), \\ \text{s.t.} \quad & -\phi_{i\mathrm{O}} y_i + Y\mu \leqslant 0, \\ & x_i - X\mu \leqslant 0, \\ & \sum_{j=1}^{N} \mu_j = 1, \\ & \mu_j \geqslant 0. \end{aligned} \tag{3.9}$$

With these concepts and terminology defined, we can now explain how to use the leave-one-out outlier detection method of Wilson (1995) relative to an inefficient frontier.

3.2. Outlier detection relative to the efficient and inefficient frontiers

One outlier detection method suggested by Wilson (1995) calculates the leave-one-out efficiency estimate to give a measure of the degree to which an observation is an outlier. While Wilson only searches for outliers relative to either an input or an output orientation, Simar (2003) suggests an observations should be distant from both an input and an output orientation in order to be an outlier. For identifying outliers relative to an efficient frontier we will heed Simar's suggestion and require the observation to be distant from both perspectives. To quantify distant, a threshold value needs to be selected. If a threshold value is chosen for one of the orientations, the reciprocal value should be used for the other orientation to specify symmetrical thresholds.

Relative to an inefficient frontier, if an observation is found to be both below this threshold value for input oriented DEA analysis (3.7) and above the reciprocal value for output oriented analysis (3.9), then the observation will be flagged as an outlier requiring further inspection. For example, if 0.5 is selected for the input oriented estimate threshold, this value corresponds to the concept the worst observation or convex combination of bad observations in the reference set excluding the observation under evaluation can produce the same level of output as the given observation using half the inputs. Similarly, 2 is the reciprocal value, if used in the output oriented analysis, this corresponds to the concept the worst observation or convex combination of bad observations in the reference set excluding the observation under evaluation can use the same level of input as the given observation and produce twice the output. Then any observation for which $\phi_{i\mathrm{I}} < 0.5$ and $\phi_{i\mathrm{O}} > 2$ would be flagged as a possible outlier. This is an example of a weak outlier threshold criterion. Of course more rigorous criteria could be selected by picking a larger (smaller) value for the input (output) oriented estimate threshold. If a large number of observations are flagged this would indicate a more rigorous threshold criteria is necessary or the quality of the peer group identified should be reevaluated. Wilson does not provide any guidance in the selection of these threshold criteria for the efficient frontier and Simar (2003) states that threshold values will be closely related to the data generation process which is specific for each group evaluated. Thus this value should be selected on a case-by-case basis.

The leave-one-out input oriented DEA inefficiency estimate is the distance of a completely inefficient observation from the inefficient frontier of the data set, not including the observation under evaluation, and can be computed using the following linear program:

$$\begin{aligned} \max_{\phi^*_{i\mathrm{I}},\mu^*_i} \quad & (\phi^*_{i\mathrm{I}}), \\ \text{s.t.} \quad & -y_i + Y^{(i)}\mu^*_i \leqslant 0, \\ & \phi^*_{i\mathrm{I}} x_i - X^{(i)}\mu^*_i \leqslant 0, \\ & \sum_{j=1}^{N} \mu^*_j = 1, \\ & \mu^*_j \geqslant 0. \end{aligned} \tag{3.10}$$

In (3.10), ϕ^*_{Ii} is the input-oriented inefficiency estimate for the ith unit, μ^*_i is a vector of intensity variables, $X^{(i)} = [x_j]\ \forall j \neq i$, $Y^{(i)} = [y_j] \forall j \neq i$, $x \in R^p_+$ denotes a $(1 \times p)$ vector of inputs, $y \in R^q_+$ denotes a $(1 \times q)$ vector of outputs, and N is the number of observations. The variables $X^{(i)}$ and $Y^{(i)}$ have dimensions $(p \times (N-1))$ and $(q \times (N-1))$,

respectively, and μ_i^* has dimensions $(1 \times (N-1))$. Similarly, the leave-one-out output oriented DEA inefficiency estimate can be calculated by the linear program

$$\begin{aligned} \min_{\phi_{iO}^*, \mu_i^*} \quad & (\phi_{iO}^*), \\ \text{s.t.} \quad & -\phi_{iO}^* y_i + Y^{(i)} \mu_i^* \leqslant 0, \\ & x_i - X^{(i)} \mu_i^* \leqslant 0, \\ & \sum_{j=1}^{N} \mu_j^* = 1, \\ & \mu_j^* \geqslant 0. \end{aligned} \tag{3.11}$$

Both (3.10) and (3.11) must be solved one time for each observation in order to develop a set of leave-one-out efficiency estimates for all observations. Observations that are candidates for outliers will have leave-one-out DEA inefficiency estimates that exceed the corresponding threshold values.

A common problem facing outlier detection methods is the masking effect. Rousseeuw and van Zomeren (1990) give a detailed discussion of this problem; in essence, the presence of an outlier hides or masks the presence of another outlier. The leave-one-out method is based on the nearest neighbor type criteria, and is particularly vulnerable to this effect. A method suggested by Simar (2003) and Wilson (1995) to lessen this problem is to apply an outlier detection process in an iterative fashion, i.e., the outlier detection method should be applied, outliers identified and removed, and the method applied again on the smaller set. This process could be applied a set number of times or until the number of outliers identified in an iteration is below a specified level.

4. Inefficient frontier: Practical implementation

As a demonstration, we use the classic Banker and Morey (1986) data set for pharmacies in the state of Iowa. There are 69 observations, 2 outputs, 3 inputs, and 1 environmental variable. The environmental variable is population and the continuous values for population are used (rather than the categorical variable constructed by Banker and Morey). For more information about the data set, see Banker and Morey (1986).

To begin, a critical value for outlier detection should be specified. The rather strict value of 1.1 was selected for the efficient frontier input oriented evaluation and the inefficient frontier output oriented evaluation. The reciprocal value of 0.91 was used for the efficient frontier output oriented evaluation and the inefficient frontier input oriented evaluation. Because the iterative method was used, the iteration on which an observation was identified as an outlier is also noted in Table 1.

Table 1
Units identified and the iteration for four outlier tests

	Efficient frontier	Inefficient frontier
1st lteration	5	17
	69	69
2nd lteration	6	15
	41	23
	55	46
3st lteration	17	47
4th lteration	4	
	7	
	12	
	44	
	53	
	65	
	67	

The two-stage bootstrapping method, algorithm 2 in Simar and Wilson (2007) was used to estimate the equation

$$\delta_i = \beta_0 + z_i \beta + \varepsilon_i, \tag{4.1}$$

where z is a (69×1) vector of the population values and δ_i is the input efficiency of unit i. The 95% bootstrap confidence interval for the parameter β based on 56 points remaining in the data set after removing outliers relative to the efficient frontier was $[-1.982, 1.712]$. Thus the result of the analysis would conclude the population has no effect on the efficiency of a pharmacy in Iowa. However, if the bootstrapping method is used on the 52 point data set with outliers removed based on both the efficient and inefficient frontier, the bootstrap confidence interval at the 95% level is $[-0.3460, -0.0005]$. This result indicates that efficiency is inversely related to population of the area in which the pharmacy is located. As this example demonstrates, misleading conclusions can be drawn from the second stage analysis if outliers are not identified and treated for both the efficient and inefficient frontiers.

5. Conclusion

This paper describes an outlier detection methodology, and formalizes the inefficient frontier. The inefficient frontier's value as an aid in outlier detec-

tion is demonstrated. Further this paper implements the iterative outlier detection method previously discussed in both Simar (2003) and Wilson (1995) and demonstrates the Simar and Wilson (2007) two-stage semi-parametric method for the Banker and Morey (1986) data set with outliers removed based only on the efficient frontier and for the data set with outliers identified based on both the efficient and inefficient frontiers. It is shown that the conclusions drawn based on the results of the two different data sets can be different and the use of outlier detection based on both the efficient and inefficient frontiers is recommended.

References

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

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

Banker, R.D., Das, S., Datar, S.M., 1989. Analysis of cost variances for management control in hospitals. Research in Governmental and Nonprofit Accounting 5.

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

Charnes, A., Cooper, W.W., Lewin, A.Y., Seiford, L.M., 1993. Data Envelopment Analysis: Theory, Methods, and Application. Quorom Books, New York.

Coelli, T., Rao, D.S.P., Battese, G.E., 1998. An Introduction to Efficiency and Productivity Analysis. Kluwer Academic Publishers, Boston.

Cook, R.D., Weisberg, S., 1982. Residuals and Influence in Regression. Chapman & Hall, New York.

Deprins, D., Simar, L., Tulkens, H., 1984. Measuring labor inefficiency in post offices. In: Machand, M., Pestieau, P., Tulkens, H. (Eds.), The Performance of Public Enterprises: Concepts and Measurements. North-Holland, Amsterdam, pp. 243–267.

Gunst, R.F., Mason, R.L., 1980. Regression Analysis and its Application. Marcel Dekker, New York.

Liu, F.-h.F., Hsu, T.-n.D., 2004. Least-efficient frontiers of data envelopment analysis-CCR model. Working Paper.

Lovell, C.A.K., 1993. Production frontiers and productive efficiency. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), The Measurement of Productive Efficiency, vol. 1. Oxford University Press, New York, pp. 3–67.

Lovell, C.A.K., 1994. Linear programming approaches to the measurement and analysis of productive efficiency. TOP 2, 175–248.

Lovell, C.A.K., Walters, L.C., Wood, L.L., 1993. Stratified models of education production using DEA and regression analysis. In: Charnes, A., Cooper, W.W., Lewin, A.Y., Seiford, L.M. (Eds.), Data Envelopment Analysis: Theory, Methods, and Application. Quorom Books, New York.

McCarty, T., Yaisawarng, S., 1993. Technical efficiency in New Jersey school districts. In: Fried, H.O., Lovell, C.A.K., Schmidt, S.S. (Eds.), The Measurement of Productive Efficiency, vol. 1. Oxford University Press, New York, pp. 271–287.

Paradi, J.C., Asmild, M., Simak, P.C., 2004. Using DEA and worst practice DEA in credit risk evaluation. Journal of Productivity Analysis 21 (2), 153–165.

Ray, S.C., 1988. Data envelopment analysis, nondiscretionary inputs and efficiency: An alternative interpretation. Socio-Economic Planning Science 22 (4), 167–176.

Ray, S.C., 1991. Resource-use efficiency in public schools: A study of Connecticut data. Management Science 37 (12), 1620–1628.

Rousseeuw, P.J., van Zomeren, B.C., 1990. Unmasking multivariate outliers and leverage points. Journal of American Statistical Association 85, 633–639.

Simar, L., 2003. Detecting outliers in frontier models: A simple approach. Journal of Productivity Analysis 20, 391–424.

Simar, L., Wilson, P.W., 2007. Estimation and inference in two-stage, semi-parametric models of production processes. Journal of Econometrics 136 (1), 31–64.

Timmer, C.P., 1971. Using a probabilistic frontier production function to measure technical efficiency. Journal of Political Economy 79, 767–794.

Wilson, P.W., 1995. Detecting influential observations in data envelopment analysis. Journal of Productivity Analysis 6, 27–45.