This paper studies the educational efficiency as determined only by the variables directly controlled by the school, isolated from the influence of other environmental characteristics, such as student’s socioeconomic status, that might influence efficiency as well. An alternative application of Simar and Wilson (2007) two-stage DEA’s approach is adopted using data from public schools in the basic education level from the Northeast Region of Brazil. The results have showed that the rank of efficiency becomes much more homogeneous after isolating from the effect of environmental variables as compared to the rank produced from a simple one-stage DEA.

Keywords: School achievement, Educational efficiency, DEA.

JEL: I21

1. INTRODUCTION

Brazil’s education system has undergone profound changes in recent years and has achieved at least one significant advance: virtually universal access to basic education. Unfortunately, this expanded attendance has not been matched by improved quality. Much to the contrary: the performance of fourth and eighth-graders on the Basic Education Development Index (Índice de Desenvolvimento da Educação Básica or IDEB) has declined since 1995 and...
although there has been some improvement since 2003, the results in 2009 still fell short of those in 1995 (MEC/INEP, 2007 and 2009).

The establishment of the National System to Evaluate Basic Education (Sistema Nacional de Avaliação da Educação Básica) in 1990, which gave rise to the current IDEB, has been a fundamental tool to monitor the quality of Brazilian education, by administering standardized achievement tests (Prova Brasil), besides providing a substantial volume of data on school practices and infrastructure, teachers’ characteristics and students’ socioeconomic background. This database is extremely valuable because it both provides data for research into the efficiency of the educational system and allows defining better public policies to address the problems identified.

A school is considered efficient if its students can achieve a maximum academic result with a minimum use of scholar resources, ceteris paribus. The problem is that the efficiency of a teaching establishment is not only affected by its resources or academic practices, it is also influenced by exogenous variables such as the student’s socioeconomic status and behavior as well as other environmental characteristics like various types of social infrastructure (provision of water, energy, sewerage, roads, safety, etc). As a result, the matter of efficiency can only be satisfactorily explained if those exogenous variables can be properly separated from the effect of the school practices and resources.

The purpose of this paper is to propose a methodology based on Data Envelopment Analysis (DEA) that addresses to this issue of efficiency using data from the public schools of the basic education level in the Northeast Region of Brazil. The article is organized into five sections including this introduction. In the next section we describe the method proposed to study school performance; in the third one we introduce the models and data utilized; in the fourth we present the initial estimations to adjust the model and the final results; and in the fifth section we make our concluding remarks.

2. USE OF DATA ENVIRONMENTAL ANALYSIS TO EVALUATE EDUCATIONAL EFFICIENCY

The literature offers basically two lines of research into school performance: econometric studies, which seek to achieve static improvements to mitigate problems of endogeneity and distinguish the school effect; and studies that apply nonparametric techniques such as Data
Envelopment Analysis to investigate questions related to school efficiency. This article proposes investigate educational efficiency by using a methodology based on Simar and Wilson (2007) work, but in an innovative fashion where the DEA estimation, differently from these authors, occurs on a second stage after cleaning out the influence of students characteristics on school efficiency.

The DEA method has been applied to various areas of knowledge, such as production engineering, management and economics. In economics alone the applications include themes such as the efficiency of agriculture production (Gomes, 2008; Souza, 2002, cited in Souza & Wilhelm, 2009), public spending (Santos et al., 2007; Sampaio de Sousa & Stosic, 2005), health services (Marinho, 2001, cited in Faria & Januzzi, 2006), the energy sector (Laurencel & Souza, 2004; Câmara, 2008; Pires, 2008), quality of life (Sousa, 2007; Bezerra & Diwan, 2001, cited in Faria & Januzzi, 2006) and education (Delgado, 2007; Delgado & Machado, 2007; Braz, 2005; Wilson, 2005; Afonso & Aubyn, 2005, Façanha & Marinho, 2001, cited in Faria & Januzzi, 2006; Silva & Fernandes, 2001, cited in Faria & Januzzi, 2006).

2.1 The DEA Model

DEA is a nonparametric technique that permits establishing the relative technical efficiency of decision making units (DMUs), each of which uses multiple inputs to produce one or more outputs. It permits identifying the efficiency of the school (DMU) based on the vector of its inputs (infrastructure, teacher qualification, etc.) available to achieve its main activity (achievement scores), permitting a varying combination of inputs to maximize efficiency, under the condition that the allocations of inputs in all the DMUs, do not consist of a single input. In this sense, it can be assumed that school administrators oriented to attain the highest efficiency will focus their choices mainly on the inputs that contribute the most to achieve its targets, and will focus less on those that have little effect on success.

In the specialized literature, the first DEA model was developed by Charnes, Cooper & Rhodes (1978), and in homage to them it is also known as the CCR model. This model consists of determining for each DMU the maximum ratio between outputs and inputs, given the available inputs of each DMU. The efficiency frontier is then constructed based only on those DMUs that have achieved the maximum output for a given level of inputs, or that have consumed a
minimum level of inputs to obtain a given amount of outputs. This construction is based on the best practices observed, since the efficiency frontier is formed from the DMUs that present, based on the observable data, the best performance in terms of the input-output ratio. Through the use of mathematical programming, the authors formulated this problem as follows:

\[
\begin{align*}
\min f_i &= \frac{\sum_{j=1}^{m} v_j x_{ji}}{\sum_{r=1}^{s} u_r y_{ri}} \\
\sum_{j=1}^{m} v_j x_{ji} &\geq 1 \\
\sum_{r=1}^{s} u_r y_{ri} &\geq 1 \\
v_j, u_r &\geq 0
\end{align*}
\] (1)

where \( f_i \) is the objective function of DMU \( i \) and \( i = 1, \ldots, n \) DMUs; \( x_{ji} \) are the inputs of the \( i \)-th DMU; \( j = 1, \ldots, m \); \( y_{ri} \) are the outputs of the \( i \)-th DMU; \( r = 1, \ldots, s \); and \( u_r \) and \( v_j \) are the weights (weighting coefficients) determined by the solution of the problem. Note that the \( i \)-th constraint in (1) expressed by the ratio \( \frac{\sum_{j=1}^{m} v_j x_{ji}}{\sum_{r=1}^{s} u_r y_{ri}} \) means that it is necessary at least a feasible input set \( X \) to produce a set of outputs \( Y \).

Equation (1) represents the fractional programming problem, which can be linearized to obtain the following linear programming model:

\[
\begin{align*}
\min g_1 &= \sum_{j=1}^{m} v_j x_{ji} \\
\sum_{r=1}^{s} u_r y_{ri} &= 1 \\
\sum_{j=1}^{m} v_j x_{ji} - \sum_{r=1}^{s} u_r y_{ri} &\geq 0 \\
v_j, u_r &\geq 0
\end{align*}
\] (2)

\^To better understand the mathematical developments, see the description provided in Charnes, Cooper & Rhodes (1978).
This is called the primal model, and for each one there is a dual formulation represented below by Equation (3). This way of composing the problem makes it simpler to resolve, by involving a smaller number of constraints (m+s < n+1), since it is advisable for the number of DMUs to be at least twice the number of variables.

\[
\begin{align*}
\text{max } f_i &= \theta \\
\text{s.t. } &\theta y_{ri} - \sum_{j=1}^{n} \lambda_{ij} y_{ji} \leq 0 \\
\sum_{j=1}^{n} \lambda_{ij} x_{ji} &\leq x_{ji} \\
\lambda_{ij}, \theta &\geq 0
\end{align*}
\]

(3)

This model seeks to find the weights, \( \lambda_{ij} \), that maximize the final output given a limited amount of inputs for each of the \( n \) schools, as well as the values of \( \theta \), which represent the schools’ efficiency indices. The values of \( \theta \) must be less than or equal to 1. When \( \theta \) is equal to one, the DMU is considered efficient. Thus, the efficiency frontier is formed by the set of points where \( \theta = 1 \). A total of \( n \) problems of this type are resolved to construct the efficiency frontier.

For each optimal solution found \((f^*)\) in each of the \( n \) objective functions, there is a \( \theta^* \) that expresses the distance of the DMU from the efficiency frontier. Hence, each inefficient DMU will have a reference for comparison \((x^*, y^*)\), on the frontier, obtained from a linear combination of the DMUs, that is, based on the best practices. \( \theta \) measures the distance of the point \((x, y)\) of the inefficient DMU to the point \((x^*, y^*)\) of the efficient DMU and hence expresses the rate of expansion of the outputs and inputs necessary for the DMU to become efficient.

The CCR model imposes the following restrictions on the technology defining the efficiency frontier: (i) the existence of constant scale returns, (ii) strong disposability of inputs and outputs, and (iii) convexity of the set of feasible combinations of inputs and outputs. However, there is nothing that guarantees the exclusive existence of constant returns associated with a given technology. It is possible, for instance, for a technology to present variable returns, increasing for a determined level of output and constant or decreasing at other levels. The model developed by Banker, Charnes & Cooper (1984), also known as the BCC model, assumes this possibility and modifies the CCR model by inserting a new constraint, \( \sum_{j=1}^{n} \lambda_{ij} = 1 \), which guarantees convexity of the combination of reference DMUs.
DEA method has two main limitations. The first is a possible inconsistency of the estimators, since because it is a nonparametric technique the convergence speed is slow and inconsistent $\theta$, can be generated. The other restriction is related to the existence of outliers that can shift the efficiency frontier too much, thus placing many DMUs in the inefficiency region when they may in reality be efficient.

To correct the problem of inconsistency of the estimators, the literature indicates the use of DEA estimation based on the bootstrap resampling scheme to eliminate bias, as proposed by Simar & Wilson (1998). The idea is based on the principle that to know the size of the bias it is necessary to obtain a distribution of the efficiency estimators, $\hat{\theta}_i$. Bootstrapping permits generating a sufficiently large series of estimates of $\hat{\theta}_i$ to obtain an empirical distribution that tends asymptotically to the true distribution of the $\theta_i$.

With respect to the problem of outliers, a common practice in DEA (e.g., Santos et al., 2007; Delgado & Machado, 2007) is to utilize a procedure developed by Sampaio de Sousa & Stosic (2005), based on jackstrap and bootstrap.

Here we treat the problem of inconsistency of the estimators through the bootstrap process proposed by Simar & Wilson (1998) and the question of identifying outlier DMUs by applying the procedure adopted by Sampaio de Sousa & Stosic (2005). It is important to remember that in correcting the bias by bootstrapping, the inconsistency biases are subtracted and for this reason the resulting values of $\theta_i$ will be less than 1.

**2.2 The two-stage model**

Most of the research done about efficiency typically applies the DEA method in two stages, where at the first stage, it is produced DEA’s efficiency estimates ($\hat{\theta}_i$) and, then, in a second stage these estimates are regressed on other exogenous variables using a parametric model, either linear ordinary least squares or censored tobit models (Aly et al 1990; Chirkos and Sears, 1994; Dietzsch and Weill, 1999; Ray, 1991; Sexton et al, 1994; Stanton, 2002, cited in Simar and Wilson, 2007).
One of the problems in a simple two-stage procedure is that the DEA’s estimates are by construction serially correlated, since a DMU is either efficient or it is related to at least another two DMUs placed on the efficient frontier (Delgado and Machado, 2007). The strategy suggested by Simar and Wilson (2007) to overcome such problem is the bootstrapping scheme to eliminate inconsistency bias. The present article also uses bootstrapping to correct for the serial correlation problem, however it proposes an alternative way to define an efficiency educational model, where the DEA technique is used in the second stage instead.

On a typical two-stage procedure the goal is to find what are the environmental characteristics, including variables that aren’t under the school control, explain the educational efficiency. Here the goal is not to find which variables explain efficiency, but instead, to provide a rank of schools that are determine only by the characteristics that are under the control of schools, independent of the socioeconomic and demographic status of students.

The idea is that the variables which affect school efficiency, but is not on the teaching establishment’s control, should be accounted for in a first stage OLS regression (4) and its residuals are used as the output variable in the second stage DEA.

\[ Y = \beta Z_i + \varepsilon_i \]  \hspace{1cm} (4)

where \( Z \) is the matrix of students academic behavior and socioeconomics and demographic status composed by the following variables: gender, age, race, mother’s years of school, father’s year of school, \( n \) of people living in the household, \( n \) of bathrooms in the household, student works, student works at home, \( n \) of computers in the household, \( n \) of books in the household, student went to kindergarten, student went to private school before, student failed before, student abandoned school before, student does math homework, student receives teacher’s compliment and parents incentive for studying. \( \varepsilon_i \) is the error term that presents the classic assumptions \( E(\varepsilon_i) = 0 \) and constant variance.

In this context, the predicted errors from (4) would capture all the variables that influence the student’s math grade but are not related to his or her socioeconomic status. The assumption is that those residuals would then incorporate all the variables associated with the school inputs (teachers, teaching resources, infrastructure, school practices and so on).
In the second stage the school efficiency is, then, estimated using the DEA method. The model and results are presented in the following section.

3. MODELS AND DATA UTILIZED TO STUDY EDUCATIONAL EFFICIENCY

The results of the math grade in Prova Brasil given to students in the fourth grade in 2007 and those of 2006 School Census compose the database used in this article to estimate the school efficiency frontier of the public schools in the Northeast region of Brazil. The School Census gathers data on schools’ physical characteristics, number of students enrolled, faculty characteristics and grade progression. In turn, the Prova Brasil measures school achievement in mathematics and Portuguese.

The total sample used to estimate the efficiency frontier was composed of 862 schools. We eliminated from the sample schools with information lacking or that presented values of zero for the school inputs considered in the model. Table 1 below shows the distribution of the sample by type of school (state or municipal).

<table>
<thead>
<tr>
<th>STATES</th>
<th>STATE SCHOOL</th>
<th>MUNICIPAL SCHOOL</th>
<th>TOTAL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alagoas</td>
<td>25</td>
<td>37</td>
<td>62</td>
</tr>
<tr>
<td>Maranhão</td>
<td>55</td>
<td>44</td>
<td>99</td>
</tr>
<tr>
<td>Paraíba</td>
<td>66</td>
<td>111</td>
<td>177</td>
</tr>
<tr>
<td>Pernambuco</td>
<td>138</td>
<td>116</td>
<td>254</td>
</tr>
<tr>
<td>Piauí</td>
<td>45</td>
<td>46</td>
<td>91</td>
</tr>
<tr>
<td>Rio Grande do Norte</td>
<td>55</td>
<td>64</td>
<td>119</td>
</tr>
<tr>
<td>Sergipe</td>
<td>32</td>
<td>28</td>
<td>60</td>
</tr>
<tr>
<td><strong>Northeast region</strong></td>
<td><strong>416</strong></td>
<td><strong>446</strong></td>
<td><strong>862</strong></td>
</tr>
</tbody>
</table>


Two different types of models were estimated. Model 1 was the two-stage DEA model described in the previous section and Model 2 was a one-stage DEA, where the output variable was the observed average math grade in each school instead of the residuals from the two-stage DEA. The variables that composed both models are described in Chart 1.
Chart 1
Input and output variables used to estimate the data envelopment analysis models

<table>
<thead>
<tr>
<th>MODEL 1</th>
<th>MODEL 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>OUTPUT VARIABLE</strong></td>
<td><strong>OUTPUT VARIABLE</strong></td>
</tr>
<tr>
<td>Average residuals from 1st stage regression (MODEL 1)</td>
<td>Average math grade of 4th graders (MODEL 2)</td>
</tr>
<tr>
<td><strong>INPUT VARIABLES</strong></td>
<td><strong>INPUT VARIABLES</strong></td>
</tr>
<tr>
<td>Total number of non-teaching staff</td>
<td>The same as MODEL 1</td>
</tr>
<tr>
<td>Total number of teachers</td>
<td></td>
</tr>
<tr>
<td>Number of classrooms</td>
<td></td>
</tr>
<tr>
<td>Percentage of teachers with teaching credentials teaching in the 1st to 4th grades</td>
<td></td>
</tr>
<tr>
<td>Number of public educational programs the school participates</td>
<td></td>
</tr>
<tr>
<td>Indicator of administrative spaces: Principal’s office, secretary’s office, teachers’ lounge</td>
<td></td>
</tr>
<tr>
<td>Indicator of building infrastructure: school building 1, included bathrooms, n° of classrooms with fans or air conditioned, drinking fountains</td>
<td></td>
</tr>
<tr>
<td>Indicator of educational spaces: library, laboratories, TV/video room, videoteca, auditorium</td>
<td></td>
</tr>
<tr>
<td>Indicator of meal facilities: kitchen, cafeteria, stove, oven, food scale</td>
<td></td>
</tr>
<tr>
<td>Indicator of basic infrastructure: public energy, water and sewerage, collected garbage</td>
<td></td>
</tr>
<tr>
<td>Indicator of teaching equipment: TV, rear projector, computers and printers</td>
<td></td>
</tr>
<tr>
<td>Indicator of sportive spaces: gymnasium, swimming pool</td>
<td></td>
</tr>
<tr>
<td>Hours students remained at school</td>
<td></td>
</tr>
<tr>
<td>Teacher’s years of school</td>
<td></td>
</tr>
<tr>
<td>Years since last graduation of the teacher</td>
<td></td>
</tr>
<tr>
<td>Teacher’s salary</td>
<td></td>
</tr>
<tr>
<td>Years teaching 4th grades</td>
<td></td>
</tr>
<tr>
<td>Teacher’s hours worked</td>
<td></td>
</tr>
<tr>
<td>Proportion taught of the prevue course content</td>
<td></td>
</tr>
<tr>
<td>Principal’s years of school</td>
<td></td>
</tr>
<tr>
<td>Years since last graduation of the Principal</td>
<td></td>
</tr>
<tr>
<td>Principal’s salary</td>
<td></td>
</tr>
<tr>
<td>Years as school Principal</td>
<td></td>
</tr>
<tr>
<td>Principal’s hours worked</td>
<td></td>
</tr>
</tbody>
</table>

1: Some schools do not have their own building, meaning that they operate in other institutions spaces, such as granted rooms from churches, gymnasium or private owners.
4. EDUCATIONAL EFFICIENCY RESULTS OF THE SCHOOLS

4.1 Initial estimates to adjust the model

After choosing the input and output variables of the models, it is necessary to define the types of returns to scale to which the educational efficiency frontier is subject. As explained in the second section, there is no consensus in the educational efficiency literature on the most suitable type of return to construct the school efficiency frontier. Many studies using DEA assume the existence of constant scale returns, where each increment in the set of inputs considered generates growth of equal magnitude on the output side. The problem with this hypothesis is that it might be true for one group of schools but not for another. It is possible that for some schools the students’ performance increases more than proportionally (increasing returns) or less than proportionally (non-increasing returns). To avoid this type of limitation, in this article we estimate the case of variable returns to scale in order to contemplate all the possible combinations. The estimation was carried out with the statistical package developed by Wilson (2006) and executed with R, called FEAR (Frontier Efficiency Analysis with R).

To minimize the problems of inconsistency of the estimators and overly influential observations (outliers), mentioned in the second section, we used two procedures. To address the problem of outliers, we used the test formulated by Sampaio de Sousa & Stosic (2005). According to this test, there were 64 outlier schools. Once the outliers were identified, they were removed from the sample and the DEA was estimated again with bias correction through bootstrap to resolve the problem of inconsistency of the estimators. This correction was done according to Simar & Wilson (1998) with 1 thousand repetitions. In this procedure, the estimated efficiency parameter of each school was subtracted in the magnitude of the bias obtained by this technique, so that the maximum efficiency value was always less than 1.

4.2 Final results

6. CONCLUSIONS
REFERENCES


MINISTÉRIO DA EDUCAÇÃO – MEC / INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA – INEP. Ensino Fundamental Regular -


WILSON, P. W. FEAR: A Software Package for Frontier Efficiency Analysis with R. December 2006. Available at: