Probabilistic Measures of Efficiency and the Influence of Contextual Variables in Nonparametric Production Models: an Application to Agricultural Research Evaluation in Brazil

Geraldo da Silva e Souza [1], Eliane Gonçalves Gomes [2] and Roberta Blass Staub [3]

[1] *geraldo.souza@embrapa.b*, [2] *eliane.gomes@embrapa.br*
Brazilian Agricultural Research Corporation (Embrapa), Parque Estação Biológica, Av. W3 Norte final, CEP 70770-901, Brasília, DF (Brazil)

[3] *roberta.blass@bcb.gov.br*
Central Bank of Brazil, Setor Bancário Sul, Quadra 3, Bloco B, Ed. Sede do BACEN, CEP 70074-900, Brasília, DF (Brazil)

Abstract
In a research institution it is important to identify which management practices have influence on the production efficiency. In this paper we assess the statistical significance of contextual variables type, size, financial resources acquisition, intensity of partnerships, processes improvements and management change. The analysis is carried out for the Brazilian Agricultural Research Corporation over the period 1999-2006. The statistical analysis uses a balanced dynamic panel data model. We conclude that only financial resources acquisition is statistically significant. The association with the production process is positive. We also found statistically significant the two lag inertial component of the ratio conditional FDH (Free Disposal Hull) to unconditional FDH indicating a two year effort to improve efficiency.

Introduction
The Brazilian Agricultural Research Corporation (Embrapa) uses a production model to monitor its research production. Embrapa has 37 research centers, spread throughout the country.

The model has multiple objectives. Firstly it allows the measurement of outputs and inputs in a systematic way. Proper qualification of inputs and outputs provides a quantitative basis that easier the understanding of the company's operations. Secondly it provides a sound basis for decision making and strategic planning at the administration level. Thirdly the computation of measures like productivity, economic efficiency and total factor productivity allows the identification of benchmarks and best procedures intended to increase overall performance and reduce differences within the organization. Finally, measures of variability in efficiency through time serve the purpose to assess the performance of the administration. In this context, the Embrapa's performance evaluation model is a decision support system.

This article is concerned with the identification of contextual variables external to the production process that may be affecting or causing efficiency. Typically these variables are in control of the institution. The assessment of their effect is of importance, since they may serve as a tuning device to promote efficiency.

The use of technical efficiency as a performance and evaluation measure raises some questions within the organization. An important one is whether or not the process generates unwanted competition among the research centers. A typical criticism is that the evaluation system may inhibit partnerships.

The identification of causal factors of efficiency demands appropriate statistical modeling. In Embrapa, Data Envelopment Analysis (DEA) technical efficiencies are computed, since 1996, under constant returns to the scale. Recently, Souza (2006) and Souza et al. (2007) assessed the influence of covariates on the DEA efficiency measurements using analysis of variance, dynamic panel data and maximum likelihood methods. A potential problem arises in this

Proceedings of ISSI 2009, pages 781–785.
Edited by B. Larsen and J. Leta.

approach: the contextual variables used may affect the production frontier. This problem is pointed out in Simar and Wilson (2007), and may affect the nature of the statistical results.
In search for an appropriate data generating mechanism for efficiency measurements and for frontier assessment, from the point of view of the influence of contextual variables, we turn to the FDH (Free Disposal Hull) measure of Deprins et al. (1984) and the extension of Daraio and Simar (2007). FDH has a probabilistic interpretation that facilitates the understanding of the production frontier, when covariates are present, via the notion of conditional probability

FDH Unconditional and Conditional Measures of Technical Efficiency

The FDH measure of technical efficiency proposed in Deprins et al. (1984) does not impose convexity on the technology set, and has an interesting probabilistic interpretation that allows the definition of a proper data generating process in the presence of contextual variables affecting the production process. The concept is defined as follows. Consider production observations (x_j, y_j), $j = 1...n$, of n producing units. The input vector x_j is a vector in R^p with nonnegative components with at least one strictly positive. The output vector y_j is a vector in R^l with nonnegative components with, at least, one strictly positive. The technical efficiency FDH of producing unit τ is taken relative to the frontier of free disposability (Free Disposal Hull) of the set (1).

$$\psi = \left\{ (x, y) \in R_+^{p+l}, y \le \sum_{j=1}^{n} \gamma_j y_j, x \ge \sum_{j=1}^{n} \gamma_j x_j, \sum_{j=1}^{n} \gamma_j = 1, \gamma_j \in \{0,1\}, j = 1...n \right\} \tag{1}$$

The input oriented FDH is given by (2) and the output oriented is given by (3).

$$\hat{\theta}(x_\tau, y_\tau) = Min\left\{ \theta; y_\tau \le \sum_{j=1}^{n} \gamma_j y_j, \theta x_\tau \le \sum_{j=1}^{n} \gamma_j x_j, \sum_{j=1}^{n} \gamma_j = 1, \gamma_j \in \{0,1\} \right\} \tag{2}$$

$$\hat{\lambda}(x_\tau, y_\tau) = Max\left\{ \lambda; \lambda y_\tau \le \sum_{j=1}^{n} \gamma_j y_j, x_\tau \ge \sum_{j=1}^{n} \gamma_j x_j, \sum_{j=1}^{n} \gamma_j = 1, \gamma_j \in \{0,1\} \right\} \tag{3}$$

A very interesting interpretation of FDH arises when the production process is described by a probability measure, defined on the product space R_+^{p+l} by random variables (X, Y). For efficiency purposes, one is interested in the probability of dominance (4).

$$H(x, y) = \Pr ob(X \le x, Y \ge y) = \Pr ob(X \le x | Y \ge y) \Pr ob(Y \ge y). \tag{4}$$

Let $F(x|y) = \Pr ob(X \le x | Y \ge y)$. The input oriented measure of technical efficiency is defined by Daraio and Simar (2007), as (5).

$$\theta(x, y) = \inf\{\theta; H(\theta x, y) > 0\} = \inf\{\theta; F(\theta x | y) > 0\}. \tag{5}$$

The empirical version is given by (6), where $I(\cdot)$ denotes an indicator function. For each producing unit in the sample this quantity is precisely the input oriented FDH measure of technical efficiency.

$$\hat{\theta}(x, y) = \frac{\sum_{j=1}^{n} I(X_j \le x, Y_j \ge y)}{\sum_{j=1}^{n} I(Y_j \ge y)} \tag{6}$$

A similar development may be considered for output orientation, leading likewise to the output oriented FDH measure of technical efficiency.
Consider now a vector Z of covariates, with values in R^k, affecting the production process. The production observations are now viewed as realizations of the conditional distribution of (X, Y) given that $Z = z$. In this case the conditional probability distribution generates the

observations. The input oriented measure of technical efficiency FDH conditional to $Z = z$ is defined by (7) and the corresponding sample estimate is (8).

$$\theta(x, y|z) = \inf\left\{\theta; H(\theta x, y|z) > 0\right\} = \inf\left\{\theta; F(\theta x|y, z) > 0\right\} \tag{7}$$

$$\hat{\theta}(x, y|z) = \frac{\sum_{j=1}^{n} I(X_j \le x, Y_j \ge y) K\left((z - z_j)/h\right)}{\sum_{j=1}^{n} I(Y_j \ge y) K\left((z - z_j)/h\right)}. \tag{8}$$

Here we assume *Z* to be absolutely continous. The function $K(\cdot)$ is a non-normal symmetric kernel concentrated in $[-1,1]^k$. The quantity h is the corresponding bandwidth for nonparametric density estimation.

In our application we use as a kernel the probability in $[-1,1]^k$ defined by the product of one-dimensional independent Epanechnikov kernels (Silverman, 1986).

For the assessment of the influence of Z in efficiency, Daraio and Simar (2007) suggested a nonparametric statistical analysis using the ratio (9) as the response variable.

$$q(x_j, y_j, z_j) = \frac{\hat{\theta}(x_j, y_j|z_j)}{\hat{\theta}(x_j, y_j)} \tag{9}$$

Here we propose a variant of this approach. For observations on a balanced panel (x_{jt}, y_{jt}, z_{jt}), $j = 1...n$, $t = 1...T$ of n producing units over T time periods we postulate (10), following Arellano and Bond (1991), Arellano and Bover (1995), Blundell and Bond (1998).

$$\begin{aligned} R_t\left(q(x_{jt}, y_{jt}, z_{jt})\right) &= c + \alpha R_{t-1}\left(q(x_{jt-1}, y_{jt-1}, z_{jt-1})\right) + \gamma R_{t-2}\left(q(x_{jt-2}, y_{jt-2}, z_{jt-2})\right) \\ &+ \sum_{f=1}^{k} \beta_f R(z_{jt}^f) + \upsilon_j + \varepsilon_{jt} \end{aligned} \tag{10}$$

The transformation $R_t(\cdot)$ denotes rank of the argument in period *t*. The quantities c, α, γ and β_f are unknown parameters, υ_j are specific random effects of the panel, the ε_{jt} iid errors with common variance σ_ε^2. The panel level effects may be correlated with the covariates. The statistical analysis is carried out using GMM methods (Greene, 2007) and is robust to the presence of serial correlation of first order in the residual structure. The use of ranks lends nonparametric properties to the analysis (Conover, 1998).

Embrapa's Production Model

The set of production variables monitored by Embrapa comprises an output y and a three dimensional input vector (x^1, x^2, x^3). The output is a weighted average of 28 production indicators. The input vector is formed by labor expenses, capital expenses and other operational expenses. For the period 1996-2006 we have balanced information on the vector (x^1, x^2, x^3, y) for all 37 Embrapa's research centers.

A proper management of the production system as a whole requires the identification of good practices and the implementation of actions with a view to improve overall performance and reduce variability in efficiency among research units. Parallel to this endeavor is the identification of non-production variables that may affect positively or negatively the system. It is of managerial interest to detect controllable attributes causing the observed best practices. In this context we consider a vector of covariates $(z_1, z_2, z_3, z_4, z_5, z_6, z_7, z_8)$ affecting the production system. Components (z_1, z_2, z_3) correspond to process improvement (mproc), financial resources acquisition (rec), and partnership (par). These are considered continuos covariates. Process improvement and intensity of partnerships are indexes. All continuos covariates are normalized by the maximum for each time. The definition of these scores can be seen in Embrapa (2006). The subvector $\left(z_4, z_5, z_6, z_7, z_8\right)$ is formed by indicator variables and corresponds to management change (adm), type and size. Two dummies are used to describe three levels for size and three levels for type, respectively. The vector of categorical variables is assumed to be exogenous to the production process.

Statistical Analysis

Table 1 shows the statistical results derived from (10). The computations were made using Matlab software kindly provided by Cinzia Daraio and Leopold Simar. The test for the presence of second autocorrelation is not significant with a p-value of 45%. The Sargan test for overidentifying restrictions does not reject the model either with a p-value of 76%. The instruments used in the analysis are first and second order differences of the response, first order differences of ranks of processes improvements, financial resources acquisitions, partnerships, the two type indicators, the two size indicators, management change indicator, and a constant term.

The effects size and type are not statistically significant with joint p-values of 84% and 86% respectively. Processes improvements, financial resources acquisition and management change have negative signs. But only financial acquisition resources is statistically significant. Therefore the response is a decreasing function of these factors. Following the interpretation of Daraio and Simar (2007), this is a case of favorable (to the production process) covariates. The intensity of partnerships is detrimental to the production process but it is not statistically significant. The lag 2 negative and statistically significant component of the response provides indication of an effort for improvement. Two periods are necessary for that to be achieved. These results are not in agreement with the analysis carried out by Souza et al. (2007), notably with respect to financial resource acquisition and management change. The differences are due more to the response used than to the statistical methods employed.

Table 1: Dynamic Panel Statistical Model. Response is rank of $q\left(x_j, y_j, z_j\right)$, the ratio of conditional to unconditional FDH measures of technical efficiency.

Variable	Coefficient	Standard Error	z	P>|z|	[95% Confidence Interval]	
Lag1	0.0377	0.2152	0.18	0.861	-0.3841	0.4595
Lag2	-0.2694	0.0905	-2.98	0.003	-0.4468	-0.0920
z1 (mproc)	-0.0108	0.0418	-0.26	0.796	-0.0928	0.0712
z2 (rec)	-0.2011	0.0977	-2.06	0.040	-0.3929	-0.0096
z3 (par)	0.0025	0.0453	0.05	0.956	-0.0863	0.0913
z4 (adm)	-0.5931	1.4980	-0.40	0.692	-3.5292	2.3429
z5 (type2)	31.7611	102.2497	0.31	0.756	-168.6446	232.1668
z6 (type3)	-83.7362	153.0349	-0.55	0.584	-383.6790	216.2067
z7 (medium)	23.7291	75.5381	0.31	0.753	-124.3228	171.7810
z8 (large)	46.7976	94.9387	0.49	0.622	-139.2788	232.8741
Intercept	32.3361	46.9948	0.69	0.491	-59.7719	124.4442

Final Considerations

The statistical assessment of the effects of contextual variables on Embrapa’s production system is carried out when the response of interest is the conditional FDH measure of technical efficiency with input orientation. The conditional FDH has an interesting probabilistic interpretation when one assumes the production model generated by a joint probability measure defined by outputs, inputs and the contextual variables. Conditioning on the absolutely continuos contextual variables, one obtains the conditional FDH. The ratio of the conditional to the unconditional FDH produces a response that can be investigated as a function of the continuous covariates and other indicator variables strictly exogenous to the production process. In this context we use a dynamic panel data model and GMM (Generalized Method of Moments) to assess the effects of contextual variables. The analysis is nonparametric. The contextual variables of interest are improvements of processes, acquisition of financial resources, management change, type and size.

We conclude that that the production process has a strong inertial component. The research centers try to improve from negative results with a two years time lag. The contextual variables processes improvements, acquisition of financial resources and management change are favorable to the production process, but only acquisition of financial resources is statistically significant. Intensity of partnerships, size and type do not show statistical significant effects.

The statistical results differ markedly from the analyses carried out with DEA measures elsewhere and the differences observed may be due to fact that CCR was used as the response variable.

Acknowledgements

To the National Council for Scientific and Technological Development (CNPq), for the financial support.

References

Arellano, M. & Bond, S. (1991). Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of Economic Studies*, 58(2), 277-297.

Arellano, M. & Bover O.(1995). Another look at the instrumental variable estimation of the error-components models. *Journal of Econometrics*, 68, 29-51.

Blundell, R. & Bond, S. (1998). Initial conditions and moment restrictions in dynamic panel-data models. *Journal of Econometrics*, 87, 115-143.

Conover, W.J. (1998). *Practical nonparametric statistics*. New York: Wiley.

Daraio, C. & Simar, L. (2007). *Advanced Robust and Nonparametric Methods in Efficiency Analysis*. New York: Springer.

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

Embrapa (2006). *Manual dos indicadores de avaliação de desempenho das unidades descentralizadas da Embrapa: Metas quantitativas - Versão para ano base 2007*. Brasília: Superintendência de Pesquisa e Desenvolvimento – Embrapa.

Greene, W.H. (2007). *Econometric Analysis*. New Jersey: Prentice Hall.

Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.

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.

Souza, G.S. (2006). Significância de efeitos técnicos na eficiência de produção da pesquisa agropecuária brasileira. *Revista Brasileira de Economia*, 60 (1), 91-117.

Souza, G.S., Gomes, E.G., Magalhães, M.C. & Avila, A.F.D. (2007). Economic efficiency of Embrapa's research centers and the influence of contextual variables. *Pesquisa Operacional*, 27, 15-26.