Chapter 9
Azorean Agriculture Efficiency by PAR

Armando B. Mendes, Veska Noncheva, and Emiliana Silva

Abstract The producers always aspire at increasing the efficiency of their production process. However, they do not always succeed in optimising their production. In the last years, the interest on Data Envelopment Analysis (DEA) as a powerful tool for measuring efficiency has increased. This is due to the large amount of data sets collected to better understand the phenomena under study and, at the same time, to the need of timely and inexpensive information.

The "Productivity Analysis with R" (PAR) framework establishes a user-friendly data envelopment analysis environment with special emphasis on variable selection, aggregation, summarisation and interpretation of the results. The starting point is the following R packages: DEA (Diaz-Martinez and Fernandez-Menendez 2008) and FEAR (Wilson 2008). The DEA package performs some models of data envelopment analysis presented in Cooper et al. (2007). FEAR is a software package for computing nonparametric efficiency estimates and testing hypotheses in frontier models. FEAR implements the bootstrap methods described in Simar and Wilson (2000).

PAR is a software framework using a portfolio of models for efficiency estimation and also providing results explanation functionality. PAR framework has been developed to distinguish between efficient and inefficient observations and

A.B. Mendes (✉)
CEEAplA, Departamento de, Matemática, Universidade dos Açores, Rua da Mãe de Deus, 9500-801 Ponta Delgada, Açores, Portugal
e-mail: amendes@uac.pt

V. Noncheva
CEEAplA, Faculty of Mathematics and Informatics, University of Plovdiv, "PaisiiHilendarski" Plovdiv, Bulgaria

E. Silva
CEEAplA, Departamento de Ciências Agrárias, Universidade dos Açores, Rua Capitão João de Ávila, 9700-042 Angra do Heroísmo, Açores, Portugal
e-mail: emiliana@uac.pt

A.B. Mendes et al. (eds.), *Efficiency Measures in the Agricultural Sector: With Applications*, DOI 10.1007/978-94-007-5739-4_9,

to explicitly advise the producers about possibilities for production optimisation. PAR framework offers several R functions for a reasonable interpretation of the data analysis results and text presentation of the obtained information. The output of an efficiency study with PAR software is self-explanatory.

We are applying PAR framework to estimate the efficiency of the agricultural system in Azores (Mendes et al. 2009). All Azorean farms will be clustered into homogeneous groups according to their efficiency measurements to define clusters of "good" practices and cluster of "less good" practices. This makes PAR appropriate to support public policies in agriculture sector in Azores.

Keywords Productivity Analysis with R • Data Envelopment Analysis • Efficiency of Azorean farms

9.1 Introduction

DEA makes it possible to identify efficient and inefficient units in a framework where results are considered in their particular context. The units to be assessed should be relatively homogeneous and were originally called Decision-Making Units (DMUs). DEA is an extreme point method and compares each DMU with only the "best" DMUs.

DEA can be a powerful tool when used wisely. A few of the characteristics that make it powerful are:

- DEA can handle multiple input and multiple output models.
- DMUs are directly compared against a peer or combination of peers.
- Inputs and outputs can have very different units. For example, one variable could be in units of lives saved and another could be in units of dollars without requiring an a priori trade-off between the two.

The same characteristics that make DEA a powerful tool can also create problems. An analyst should keep these limitations in mind when choosing whether or not to use DEA:

- Since DEA is an extreme point technique, noise such as measurement error can cause significant problems.
- DEA is good at estimating "relative" efficiency of a DMU, but it converges very slowly to "absolute" efficiency. In other words, it can tell you how well you are doing compared to your peers but not compared to a "theoretical maximum".

PAR combines DEA with different statistical methods. DEA is applied to distinguish between efficient and inefficient observations of performances. Different statistical methods are applied to assist DEA. For example, canonical correlation analysis assists DEA with both variable aggregation and variable selection. PAR methodology is implemented in R. The output of the PAR computer program is self-explanatory.

At first, we will define the performance of a farm. A natural measure of performance is a productivity ratio: the ratio of outputs to inputs, where larger values of this ratio are associated with better performance. Performance is a relative concept. For example, the performance of the meat farm in 2008 could be measured relative to its 2007 performance or it could be measured relative to the performance of another farm in 2008. This farm can also analyse the relative performance of units within the farm.

9.2 PAR: A Tool for Measuring Efficiency of Azorean Farms

9.2.1 Basic Term Definitions

We are going to provide some informal definitions of the following terms.

9.2.1.1 Productivity

Productivity can be simply defined as the ratio between outputs and inputs of an economic system. When we refer to productivity, we are referring to total farm productivity, which is a productivity measure involving all factors of production (all inputs and all outputs). The land productivity yields in farming are a partial measure of productivity. The partial productivity measures can provide a misleading indication of overall productivity when considered in isolation.

9.2.1.2 Production Frontier Line

The production frontier line may be used to define the relationship between the input and output. The production frontier represents the maximum output attainable from each input level. It reflects the current state of technology in the farm. Farms operate either on that frontier, if they are technically efficient, or beneath the frontier, if they are technically inefficient.

Efficiency frontier represents a standard of performance that the firms not on the frontier could try to achieve. Firms on the frontier are 100% efficient.

Note that this does not mean that the performance of the DMUs on the efficiency frontier cannot be improved. It may or may not be possible. However, the available data does not give any idea on the extent to which their performance can be improved.

The DMUs on the efficiency frontier are the best DMUs with the data that we have. As we do not have another DMU having better performance, we should assume that these are the best achievable performances. We rate the performance of all other firms in relation to this best achieved performance. Thus, we are talking of only relative efficiencies, not absolute efficiencies.

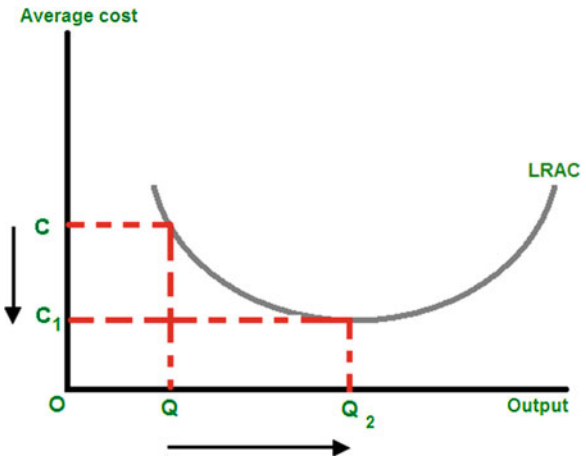

Fig. 9.1 Increase in output from Q to Q_2 causes a decrease in the average cost of each unit from C to C_1

Such an analysis, using efficiency frontier, is often termed as frontier analysis. This efficiency frontier forms the basis of the efficiency analysis. The efficiency frontier envelops the available data, hence the name Data Envelopment Analysis (DEA).

Consider the DMU which does not lie on the frontier. This DMU is inefficient. The following question arises: Can we make a quantitative estimate of its efficiency in relation to the performance of the best firm lying on the frontier?

9.2.1.3 Economies of Scale (ES)

The increase in efficiency of production as the number of goods being produced increases is known as economies of scale. Typically, an agricultural company that achieves economies of scale lowers the average cost per unit through increased production since fixed costs are shared over an increased number of goods.

Economies of scale means that as a company grows and production units increase, the company will have a better chance to decrease its costs. In Fig. 9.1, a generic Long-Run Average Cost (LRAC) curve is represented to illustrate the concept.

Economies of scale are the cost advantages that a firm obtains due to expansion. This should not be confused with increasing returns to scales where simply increasing output within current capacity reduces the short-run cost per unit.

Figure 9.1 shows a simple example, and in real life, there are countering forces of diseconomies of scale. Diseconomies of Scale (DS) are the forces that cause larger firms to produce goods and services at increased per unit costs. As these forces balance, an optimum production volume can be found referred to as constant returns to scales.

Economies of scale refers to the decreased per unit cost as output increases. More clearly, the initial investment of capital is spread over an increasing number of units of output, and therefore, the marginal cost of producing a good or service decreases as production increases (note that this is only in an industry that is experiencing economies of scale).

As we mentioned before, diseconomies may also occur. They could stem from inefficient managerial or labour policies, over-hiring or deteriorating transportation networks (external DS). Furthermore, as a company's scope increases, it may have to distribute its goods and services in progressively more dispersed areas. This can actually increase average costs resulting in diseconomies of scale.

Some efficiencies and inefficiencies are more location specific, while others are not affected by area. If a company has many plants throughout a country, they can all benefit from costly inputs such as advertising. However, efficiencies and inefficiencies can alternatively stem from a particular location, such as a good or bad climate for farming. When ES or DS are location specific, trade is used in order to gain access to the efficiencies.

The key to understanding economies of scale and diseconomies of scale is that the sources vary. A company needs to determine the net effect of its decisions affecting its efficiency and not just focus on one particular source. Thus, while a decision to increase its scale of operations may result in decreasing the average cost of inputs (volume discounts), it could also give rise to diseconomies of scale if its subsequently widened distribution network is inefficient because not enough transport trucks were invested in as well. Thus, when making a strategic decision to expand, companies need to balance the effects of different sources of economies of scale and diseconomies of scale so that the average cost of all decisions made is lower, resulting in greater efficiency all around.

9.2.1.4 Returns to Scales

Refers to a technical property of production that examines changes in output subsequent to a proportional change in all inputs (where all inputs increase by a constant factor). If output increases by that same proportional change, then there are constant returns to scales (CRS). If output increases by less than that proportional change, there are decreasing returns to scales (DRS). If output increases by more than that proportion, there are increasing returns to scales (IRS).

As a short example, where all inputs increase by a factor of 2, new values for output should be:

- Twice the previous output given = a constant returns to scales (CRS)
- Less than twice the previous output given = a decreased returns to scales (DRS)
- More than twice the previous output given = an increased returns to scales (IRS)

9.2.1.5 Allocative Efficiency

Allocative efficiency is a situation in which the limited resources of a firm are allocated in accordance with the wishes of consumers. An allocatively efficient economy produces an "optimal mix" of commodities.

A firm is allocatively efficient when its price is equal to its marginal costs in a perfect market.

Allocative efficiency means efficient distribution of resources: an economic situation where no possible reorganisation of production resources can make some consumers better off without making other consumers worse off.

If price information is available and a behaviour objective is appropriate, then it is possible to measure allocative efficiencies as well as technical efficiencies. Behaviour objectives could be cost minimisation or revenue or profit maximisation. Cost minimisation and revenue maximisation together imply profit maximisation.

9.2.1.6 Factors Which Could Influence the Efficiency of a Farm

These factors are not traditional inputs and are assumed not under the control of the manager. Some examples are:

- Ownership differences (public/private, corporate/noncorporate)
- Coal-fired electric power station influenced by coal quality
- Electric power distribution networks influenced by population density and average customer size
- Schools influenced by socio-economic status of children and city/country location
- Labour union power
- Government regulations

9.2.2 DEA Models

As we mentioned above, the organisational units and farms are more generally called Decision-Making Units (DMUs). DMUs can also be manufacturing units, departments of a big organisation such as universities, schools, bank branches, hospitals, medical practitioners, power plants, police stations, tax offices, prisons, defence bases or a set of firms. In the area of tourism, DMUs can be hotels, motels, destinations, tourism websites and so on.

Efficiency of a decision-making unit is defined as the ratio between a weighted sum of its outputs and a weighted sum of its inputs. We can find the DMU (or the DMUs) having the highest ratio. We call it DMU_o. Then we can compare the performance of all other DMUs relative to the performance of DMU_o. We can calculate the relative efficiency of the DMUs.

Suppose there are n DMUs, DMU_j, $j = 1, 2, \ldots, n$. Suppose m input items and s output items are selected:

- Let the input data for DMUs be $\underset{m\times n}{X} = \left(x_{ij}\right)_{i=1,\ldots,m;\, j=1,\ldots,n}$.
- Let the output data for DMUs be $\underset{s\times n}{Y} = \left(y_{kj}\right)_{k=1,\ldots,s;\, j=1,\ldots,n}$.

Given the data, we can measure the efficiency of each DMU_j, $j = 1, 2, \ldots, n$. Hence, we need n optimisations (one for each DMU to be evaluated).

Let the DMU we are evaluating be designated as DMU_o ($o = 1, 2, \ldots, n$).

9.2.2.1 Charnes, Cooper and Rhodes (CCR) Model

We will define the CCR-efficiency taking into account all input excesses and output shortfalls. The input-oriented CCR model aims to minimise inputs while satisfying at least the given output levels. The output-oriented CCR model attempts to maximise outputs without requiring more of any of the observed input variables.

Based on the matrix (X, Y), where X is an $(m \times n)$ matrix and Y is an $(s \times n)$ matrix, the envelopment form of the CCR model is expressed as follows:

$$\min_{\theta,\lambda} \theta \tag{9.1}$$

subject to $\theta x_o - X\lambda \geq 0$, $Y\lambda \geq y_o$ and $\lambda \geq 0$ where, for any DMU_o, $\underset{m\times 1}{x_o} = (x_{1o}, x_{2o}, \ldots, x_{mo})^{\mathrm{T}}$, θ is a real variable and $\underset{n\times 1}{\lambda} = (\lambda_1, \ldots, \lambda_n)^{\mathrm{T}}$ is a non-negative vector.

For all DMUs, together we have the following matrix notations:

$$\underset{n\times 1}{\theta}, \underset{n\times n}{\lambda} = \left(\lambda_{jj}\right)_{j=1,\ldots,n;\, j=1,\ldots,n} \text{and} \min_{\theta,\lambda} \underset{1\times n}{\theta} \tag{9.2}$$

subject to $\underset{m\times 1}{x_o}\ \underset{1\times n}{\theta} - \underset{m\times n}{X}\ \underset{n\times n}{\lambda} \geq \underset{m\times n}{0}$, $\underset{s\times n}{Y}\ \underset{n\times 1}{\lambda} \geq \underset{s\times 1}{y_o}$ and $\underset{n\times 1}{\lambda} \geq \underset{n\times 1}{0}$

The optimal θ is denoted by θ^*. It is greater than zero and not greater than 1, or $0 < \theta^* \leq 1$.

We define slack vectors by $\underset{m\times 1}{s^-} = \underset{m\times 1}{x_o}\ \underset{1\times 1}{\theta} - \underset{m\times n}{X}\ \underset{n\times 1}{\lambda}$ and $\underset{s\times 1}{s^+} = \underset{s\times n}{Y}\ \underset{n\times 1}{\lambda} - \underset{s\times 1}{y_o}$.

Definition (CCR-efficiency): If an optimal solution (θ^*, λ^*, s^{-*}, s^{+*}) of the CCR model satisfies $\theta* = 1$, $s^{-*} = 0$ and $s^{+*} = 0$, then the DMU_o is called CCR-efficient. Otherwise, the DMU_o is called CCR-inefficient.

The condition $\theta^* = 1$ is referred to as "radial efficiency". The term "weak efficiency" is sometimes used when attention is restricted to the condition $\theta^* = 1$ (also called "Farrell efficiency"). The conditions $\theta^* = 1$, $s^{-*} = 0$ and $s^{+*} = 0$, taken together, describe what is also called "Pareto-Koopmans" or "strong" efficiency.

Definition (Pareto-Koopmans efficiency): A DMU is fully efficient if and only if it is not possible to improve any input or output without worsening some other input or output.

Definition (reference set): For an inefficient DMU_o, we define its reference set E_o by $E_o = \{j \mid \lambda j^* > 0\}, j = 1, \ldots, n.$

An optimal solution can be expressed as

$$\begin{aligned} \underset{m\times 1}{x_o}\ \underset{1\times 1}{\theta^*} &= \underset{m\times n}{X}\ \underset{n\times 1}{\lambda^*} + \underset{m\times 1}{s^{-*}} = \sum_{j\in E_o} x_j \lambda_j^* + s^{-*} \\ \underset{s\times 1}{y_o} &= \underset{s\times n}{Y}\ \underset{n\times 1}{\lambda^*} - \underset{s\times 1}{s^{+*}} = \sum_{j\in E_o} y_j \lambda_j^* - s^{+*} \end{aligned} \tag{9.3}$$

The efficiency of (x_o, y_o) for DMU_o can be improved by the formula

$$\begin{aligned} \underset{m\times 1}{\hat{x}_o} &= \underset{m\times 1}{x_o}\ \underset{1\times 1}{\theta^*} - \underset{m\times 1}{s^{-*}} \le \underset{m\times 1}{x_o} \\ \underset{s\times 1}{\hat{y}_o} &= \underset{s\times 1}{y_o} + \underset{s\times 1}{s^{+*}} \ge \underset{s\times 1}{y_o} \end{aligned} \tag{9.4}$$

This formula for improvement is called the CCR-projection.

Theorem: *The improved activity* $(\hat{x}_o, \hat{y}_o)$ *defined by the CCR-projection is CCR-efficient.*

Corollary to theorem: *The point with coordinates* $(\hat{x}_o, \hat{y}_o)$

$$\begin{aligned} \underset{m\times 1}{\hat{x}_o} &= \underset{m\times 1}{x_o}\ \underset{1\times 1}{\theta^*} - \underset{m\times 1}{s^{-*}} = \sum_{j\in E_o} x_j \lambda_j^* \\ \underset{s\times 1}{\hat{y}_o} &= \underset{s\times 1}{y_o} + \underset{s\times 1}{s^{+*}} = \sum_{j\in E_o} y_j \lambda_j^* \end{aligned} \tag{9.5}$$

is the point on the efficient frontier used to evaluate the performance of DMU_o.

9.2.2.2 The Output-Oriented CCR Model

The output-oriented CCR model attempts to maximise outputs while using no more than the observed amount of any input.

The slack (t^-, t^+) of the output-oriented model is defined by

$$\begin{aligned} t^- &= x_o - X\mu \\ t^+ &= Y\mu - \eta y_o \end{aligned} \tag{9.6}$$

η^* satisfies $\eta^* \geq 1$. The higher the value of η^*, the less efficient the DMU is. η^* expresses the output enlargement rate.

An input-oriented CCR model is efficient for any DMU if and only if it is also efficient when the output-oriented CCR model is used to evaluate its performance. The solution of the output-oriented CCR model may be obtained from that of the input-oriented CCR model.

The improvement using output-oriented CCR model is expressed by

$$\begin{aligned} \hat{x_o} &= x_o - t^{-*} = \sum_{j \in E_o} x_j \mu_j^* \\ \hat{y_o} &= \eta^* y_o + t^{+*} = \sum_{j \in E_o} y_j \eta_j^* \end{aligned} \tag{9.7}$$

9.2.2.3 Banker, Charnes and Cooper (BCC) Model

The BCC problem is solved using a two-phase procedure. In the first phase, we minimise θ_B, and, in the second phase, we maximise the sum of the input excesses and output shortfalls, keeping $\theta_B = \theta_B{}^*$. Here $\theta_B{}^*$ is the optimal value obtained in the first phase. An optimal BCC solution is represented by $(\theta_B{}^*, \lambda^*, s^{-*}, s^{+*})$, where s^{-*} and s^{+*} represent the maximal input excesses and output shortfalls, respectively.

Definition (BCC-efficiency): If an optimal BCC solution $(\theta_B{}^*, \lambda^*, s^{-*}, s^{+*})$ satisfies $\theta_B{}^* = 1$, $s^{-*} = 0$ and $s^{+*} = 0$, then the DMU$_o$ is called BCC-efficient.

We have the following formula for improvement:

$$\hat{x_o} = \theta_B * x_o - s^{-*} - *, \hat{y_o} = y_o + s^{+*} \tag{9.8}$$

Theorem: *The improved activity* $(\hat{x_o}, \hat{y_o})$ *is BCC-efficient.*

Theorem: *A DMU that has a minimum input value for any input item, or a maximum output value for any output item, is BCC-efficient.*

9.2.2.4 The Increasing Returns to Scales Model (IRS) and the Decreasing Returns to Scales Model (DRS) or Relaxation of the Convexity Condition

The BCC envelopment model can be extended by relaxing the convexity condition $e\lambda = 1$ to $L \leq e\lambda \leq U$, where L, $(0 \leq L \leq 1)$ and U, $(1 \leq U)$ are lower and upper bounds for the sum of the λ_j. Notice that $L = 0$, $U = \infty$ corresponds to the CCR model and $L = U = 1$ corresponds to the BCC model. Two typical extensions are discussed below.

The case $L = 1$, $U = \infty$ is called the *Increasing Returns to Scales* (IRS) or Non-Decreasing Returns to Scales (NDRS) model. The case $L = 0$, $U = 1$ is called the *Decreasing Returns to Scales* (DRS) or the *Non-Increasing Returns to Scales* (NIRS) model.

9.2.2.5 The Increasing Returns to Scales Model (IRS)

The constraint on λ is $e\lambda \geq 1$. The interpretation of this constraint is that we cannot reduce the scale of DMU but it is possible to expand the scale to infinity. The output/input ratio for any point on the efficient frontier is not decreasing with respect to input. The term NDRS is derived from that fact. That is, a proportional increase in output is always at least as great as the related proportional increase in output and is always at least as great as the related proportional increase in input. In mathematical terms, $\Delta y / y \geq \Delta x / x$, where Δy, Δx are the increases to be made from a frontier point with coordinate (x, y). *This model focuses on the scale efficiencies of relatively small DMUs.*

9.2.2.6 The Decreasing Returns to Scales (DRS) Model

The constraints on λ are $0 \leq e\lambda \leq 1$. The interpretation of these constraints is that scaling up of DMUs is interdicted and scaling down is permitted. The output/input ratio of efficient frontier points is decreasing with respect to the input scale. That is, $\Delta y / y = \Delta x / x$ for the first segment on the frontier and strict inequality $\Delta y / y < \Delta x / x$ holding thereafter. This model puts the emphasis on larger DMUs where returns to scales is decreasing.

It is logically true that for every DMU we have the relations $\theta^*_{\text{CCR}} \leq \theta^*_{\text{IRS}}$, $\theta^*_{\text{DRS}} \leq \theta^*_{\text{BCC}}$.

9.2.2.7 Model Sources of Inefficiency

It is interesting to investigate the sources of inefficiency that a DMU might have. Are they caused by the inefficient operation of the DMU itself or by the disadvantageous conditions under which the DMU is operating?

For this purpose, comparisons of the (input-oriented) CCR and BCC scores deserve consideration. The CCR model assumes the constant returns to scales production possibility set. It is postulated that the radial expansion and reduction of all observed DMUs and their non-negative combinations are possible and hence the CCR score is called global technical efficiency. The BCC model assumes that convex combinations of the observed DMUs form the production possibility set and the BCC score is called local pure technical efficiency. If a DMU is fully efficient in both the CCR and BCC scores, it is operating in the most productive scale size.

If a DMU has full BCC-efficiency but a low CCR score, then it is operating locally efficiently but not globally efficiently due to the scale size of DMU. Thus,

it is reasonable to characterise the scale efficiency of a DMU by the ration of CCR and BCC scores. We define scale efficiency as follows:

Definition: Let the CCR and BCC scores of a DMU be θ^*_{CCR} and θ^*_{BCC}, respectively. The scale efficiency (SCAL) is defined by

$$\text{SE} = \frac{\theta^*_{\text{CCR}}}{\theta^*_{\text{BCC}}} \tag{9.9}$$

SCAL is not greater than 1. The BCC score expresses the (local) Pure Technical Efficiency (PTE) under variable returns to scales circumstances. The CCR score is called the (global) Technical Efficiency (TE) since it takes no account of scale effect as distinguished from PTE. For a BCC-efficient DMU with constant returns to scales characteristics (i.e. in the most productive scale size), the scale efficiency (SCAL) is 1.

9.2.2.8 SBM Model

We introduce a new measure ρ called SBM (Slacks-Based Measure). It is invariant to the units of measure used for the different inputs and outputs. This new measure is a scalar that yields the same efficiency value when distances are measured in either kilometres or miles. More generally, this measure is the same when x_{io} and x_{ij} are replaced by $k_i x_{io} = \hat{x}_{io}$ and $k_i x_{ij} = \hat{x}_{ij}$ and y_{ro} and y_{rj} are replaced by $c_r y_{ro} = \hat{y}_{ro}$ and $c_r y_{rj} = \hat{y}_{rj}$, where k_i and c_r are positive constants, $i = 1, \ldots, m$, $r = 1, \ldots, s$. This property is known as "units invariant". The SBM measure is monotone decreasing in each input and output slack. This property is known as "monotone".

Slacks-based measure ρ can be interpreted as the ratio of mean input and output mix inefficiencies.

Theorem: *If DMU* ***A*** *dominates DMU* ***B*** *so that* $x_A \le x_B$ *and* $y_A \le y_B$, *then* $\rho^*_A \ge \rho^*_B$.

Definition (SBM-efficient): A DMU (x_o, y_o) is SBM-efficient if and only if $\rho^* = 1$.

This condition is equivalent to $s^{-*} = 0$ and $s^{+*} = 0$, i.e. no input excess and no output shortfall in an optimal solution.

For an SBM-inefficient DMU (x_o, y_o), we have the expression:

$$\begin{aligned} x_o &= X\lambda^* + s^{-*}, \\ y_o &= Y\lambda^* - s^{+*}. \end{aligned} \tag{9.10}$$

The DMU (x_o, y_o) can be improved and becomes efficient by deleting the input excesses and augmenting the output shortfalls. This is accomplished by SBM-projection expressed by the following formulae, called SBM-projection:

$$\begin{aligned} \hat{x}_o &= x_o - s^{-*}, \\ \hat{y}_o &= y_o + s^{+*}. \end{aligned} \tag{9.11}$$

which are the same as for the additive model.

We will define the reference set for (x_o, y_o) as the following:

Definition (reference set): The set of indices R_o corresponding to positive $\lambda_j{}^*s$ is called the reference set for (x_o, y_o).

Using the reference set R_o, we can express $(\hat{x_o}, \hat{y_o})$ by

$$\hat{x_o} = \sum_{j \in R_o} x_j \lambda_j^*$$
$$\hat{y_o} = \sum_{j \in R_o} y_j \lambda_j^* \tag{9.12}$$

This means that the point on the efficient frontier $(\hat{x_o}, \hat{y_o})$ is expressed as a positive combination of the members of the reference set R_o. The members of the reference set R_o are also efficient.

Theorem: *The optimal SMB ρ^* is not greater than the optimal CCR θ^*.*

This theorem reflects the fact that SBM accounts for *all inefficiencies* whereas θ^* accounts only for *"purely technical" inefficiencies*.

The relation between CCR-efficiency and SMB-efficiency is given in the following theorem:

Theorem: *A DMU (x_o, y_o) is CCR-efficient if and only if it is SMB-efficient.*

9.2.2.9 Outlier Detection in PAR

The main drawback of deterministic frontier models is that they are very sensitive to outliers and extreme values and that noisy data are not allowed. We perform outlier analysis using the method described in Wilson (1993). This chapter describes an influence-function approach for detecting outliers in the context of frontier models.

The graphic analysis based on outlier statistic developed in Wilson (1993) and implemented in FEAR is used to identify observations in DEA models that are possible outliers. A line in the log-ratio plot connects the second smallest value of the ratios for each observation deleted to illustrate the separation between the smallest ratios for each observation. The plot is approximately linear under the homogeneity model. Under the heterogeneity model, the log-ratio plot shows convexity.

9.2.2.10 Some Notes on CCA and Some Related Methods

Canonical correlation analysis (CCA) is a multidimensional exploratory statistical method.

A canonical correlation is the correlation of two latent (canonical) variables, one representing a set of independent variables, the other a set of dependent variables.

Each set may be considered a latent variable based on measured original variables in its set. The canonical correlation is optimised such that the linear correlation between the two latent variables (called canonical variates) is maximised. There may be more canonical variates relating the two sets of variables. The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables. For each canonical variate, we can also assess how strongly it is related to measured variables in its own set or the set for the other canonical variate.

Both methods, principal components analysis (PCA) and CCA, have the same mathematical background. The main purpose of CCA is the exploration of sample correlations between two sets of quantitative variables, whereas PCA deals with one data set in order to reduce dimensionality through linear combination of initial variables.

Another well-known method can deal with the same kind of data: Partial Least Squares (PLS) regression. However, the object of PLS regression is to explain one or several response variables (outputs) in one set by way of variables in the other one (the input). On the other hand, the object of CCA is to explore correlations between two sets of variables whose roles in the analysis are strictly symmetric. As a consequence, mathematical principles of both PLS and CCA methods are fairly different.

9.2.2.11 Variable Aggregation in PAR

The question of obtaining an appropriate aggregate input from appropriate individual inputs is an important one. A natural way to define an aggregate input is to assume a linear structure of aggregation of the input variables. One of the most important issues here is the choice of weights in the aggregation.

A natural extension of the aggregation of input or output techniques is the use of weight restrictions. The use of weight restrictions is a much more subtle technique. For example, instead of eliminating an unimportant input or output, which is the same as assigning a zero weight to it, we may restrict its weight to be low in relation to the more important inputs and outputs. This way the unimportant parameter will still count in the overall model but only up to the specified limit of "importance".

Weight choice may be done by the researcher according to his opinion about the contribution of each variable. In our approach, we use Canonical Correlation Analysis (CCA) to aggregate automatically both input and output data sets.

Obviously the input and output sets of variables in a production process are related. We are concerned with determining a relationship between the two sets of variables. The aim is the linear combinations that maximise the canonical correlation to be found. Such a linear combination is called *canonical variate*.

In this chapter, we propose CCA to aggregate both input and output variables to get final input and output, respectively.

The aggregation in PAR approach is not fixed, and because of it, we are giving the answer of the following two important questions that arise frequently.

9.2.2.12 Variable Selection in PAR

Variable selection in DEA is problematic. The estimated efficiency for any DMU depends on the inputs and outputs included in the model. It also depends on the number of outputs plus inputs. It is clearly important to select parsimonious specifications and to avoid as far as possible models that assign full high efficiency ratings to DMUs that operate in unusual ways.

In practice, when we apply DEA, the number of DMUs should be greater than the total amount of variables in both sets. Usually in real-world applications, the number of DMUs is restricted. Because of it, one of the most important steps in the modelling using DEA is the choice of input and output variables.

Variable selection is crucial to the process as the omission of some of the inputs can have a large effect on the measure of efficiency. It is now recognised that improper variable selection often results in biased DEA evaluation results.

The attention to variable selection is particularly crucial since the greater the number of input and output variable, the less discerning are the DEA results (Jenkins and Anderson 2003). However, there is no consensus on how best to limit the number of variables.

Several methods have been proposed that involve the analysis of correlation among the variables, with the goal of choosing a set of variables that are not highly correlated with one another. Unfortunately, studies have shown that these approaches yield results which are often inconsistent in the sense that removing variables that are highly correlated with others can still have a large effect on the DEA results (see Nunamaker 1985). Other approaches look at the change in the efficiencies themselves as variables are added and removed from the DEA models, often with a focus on determining when the changes in the efficiencies can be considered statistically significant. As part of these approaches, procedures for the selection of variables to be included in the model have been developed by sequentially applying statistical techniques.

Another commonly used approach for reducing the list of variables for inclusion in the DEA model is to apply regression and correlation analysis (Lewin et al. 1982). This approach purports those variables which are highly correlated with existing model variables are merely redundant and should be omitted from further analysis. Therefore, a parsimonious model typically shows generally low correlations among the input and output variables, respectively, Chilingerian (1995) and Salinas-Jimenez and Smith (1996).

The authors Norman and Stoker (1991) noted that the observation of high statistical correlation alone was not sufficient. A logical causal relationship to explain why the variable influenced performance was necessary. Another application of variable selection based on correlating the efficiency scores can be found in Sigala et al. (2004).

In this chapter, we propose CCA to be used in order for the most appropriate variables to be selected. In PAR approach, we apply CCA to select both input and output variables and to get final input and output sets, respectively.

9.3 Azorean Farms' Efficiency Measurement

The Azores islands belong to the Portuguese territory with a population of about 250,000 inhabitants. The main economic activity is dairy and meat farming. Dairy policy depends on Common Agricultural Policy of the European Union and is limited by quotas. In this context, decision makers need knowledge for deciding the best policies in promoting quality and best practices. One of the goals of our work is to provide Azorean government with a reliable tool for measurement of productive efficiency of the farms.

The names of all input variables used in analysis are the following: EquipmentRepair, Oil, Lubricant, EquipmentAmortization, AnimalConcentrate, VeterinaryAndMedicine, OtherAnimalCosts, PlantsSeeds, Fertilizers, Herbicides, LandRent, Insurance, MilkSubsidy, MaizeSubsidy, SubsidyPOSEIMA, AreaDimension and DairyCows. The names of output variables are Milk and Cattle.

We start the data analysis with outlier detection. One outlier obtained in Terceira data was the result of a recording error and was corrected. We used again the statistical methodology presented in Wilson (1993) and implemented it in FEAR package to look for new atypical observations. Using the graphical analysis presented in Fig. 9.2, another observation could also be identified as an outlier. However, data from Terceira Island is viewed as coming from a probability distribution, and it is quite possible to observe one point with low probability. One would not expect to observe many such points, given their low probability. The fact that a particular observation has low probability of occurrence is not sufficient to warrant the conclusion that this observation is an error. More errors in the available data are not identified.

Canonical correlation analysis aims at highlighting correlations between input and output data sets. Two preliminary steps calculate the sample correlation coeffi-

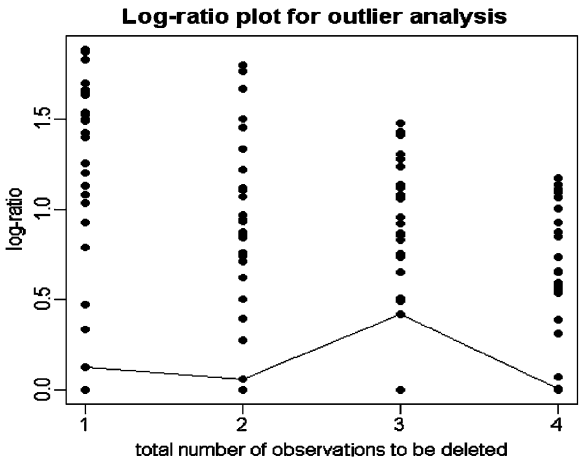

Fig. 9.2 Plot produced by the outlier detection procedure

Table 9.1 Sample correlation coefficients

	Milk	Cattle
EquipmentRepair	0.399089550	0.449336923
Oil	0.349190515	−0.023206764
Lubricant	0.009272362	−0.171455723
EquipmentAmortization	0.051043354	−0.077088336
AnimalConcentrate	0.914685924	0.537983929
VeterinaryAndMedicine	0.707943660	0.370392398
OtherAnimalCosts	0.724266952	0.407358115
PlantsSeeds	0.719946680	0.304399253
Fertilizers	0.781448807	0.452145566
Herbicides	0.497643020	0.347245965
LandRent	0.722516988	0.343699321
Insurance	−0.072519332	0.002379461
MilkSubsidy	0.746508776	0.431464776
MaizeSubsidy	0.751413121	0.526768325
SubsidyPOSEIMA	0.724407535	0.083726114
AreaDimension	0.536678292	0.279164537
DairyCows	0.776032879	0.348513730

cients and visualise the correlation matrixes. All sample correlation coefficients are presented in Table 9.1.

This table highlights a significant correlation between Milk and AnimalConcentrate and nearly null correlation between Milk and Lubricant, Milk and EquipmentAmortization and Milk and Insurance.

In practice, the number of DMUs should be greater than the total amount of variables in both input and output sets. Any resource used by an Azorean dairy farm is treated as an input variable, and because of it, the list of variables that provide an accurate description of the milk and meat production process is large.

This example is focused on measuring efficiency when the number of DMUs is few and the number of explanatory variables needed to compute the measure of efficiency is too large. We approach this problem from a statistical standpoint through both variable selection and variable aggregation approaches.

The results from CCA are printed in the following table.

From Table 9.2, we can conclude that both canonical variates are predominantly associated with the original inputs AnimalConcentrate and Fertilizers and with the original output variable Milk. In this way, we select the two input variables AnimalConcentrate and Fertilizers and one output variable Milk.

On Fig. 9.3, the input and output variables are plotted on the first two canonical variates. Variables with a strong relation are projected in the same direction from the origin. The greater the distance from the origin, the stronger the relation is. The following variables, AnimalConcentrate, VeterinaryAndMedicine, OtherAnimalCosts, MilkSubsidy, MaizeSubsidy, Herbicides, Fertilizers, PlantsSeeds, LandRent, AreaDimension, DairyCows and Milk, are a set of variables with a stronger relation than the rest. In this set, AnimalConcentrate, DairyCows, VeterinaryAndMedicine, OtherAnimalCosts and MilkSubsidy are the variables with the most strong relation. MaizeSubsidy and Herbicides are also variables with a strong relation.

Table 9.2 Correlations of the original outputs with both aggregated input and output

	$scores$corr.*Y*.*x*scores	$scores$corr.*Y*.*y*scores
Milk	−0.9529591	−0.9953781
Cattle	−0.5225409	−0.5458007
	$scores$corr.*X*.*x*scores	$scores$corr.*X*.*y*scores
EquipmentRepair	−0.44487248	−0.42591381
Oil	−0.34213524	−0.32755482
Lubricant	0.01024649	0.00980983
EquipmentAmortization	−0.04167289	−0.03989696
AnimalConcentrate	−0.96395974	−0.92287966
VeterinaryAndMedicine	−0.74087590	−0.70930276
OtherAnimalCosts	−0.76117503	−0.72873682
PlantsSeeds	−0.74525915	−0.71349921
Fertilizers	−0.82269954	−0.78763940
Herbicides	−0.53062365	−0.50801061
LandRent	−0.75224389	−0.72018629
Insurance	0.07133021	0.06829041
MilkSubsidy	−0.78586254	−0.75237225
MaizeSubsidy	−0.80148885	−0.76733263
SubsidyPOSEIMA	−0.72469294	−0.69380945
AreaDimension	−0.56145996	−0.53753280
DairyCows	−0.80562574	−0.77129323

Both the original inputs and outputs are aggregated into overall measures called aggregate input variate and aggregate output variate.

Then we use aggregated input and output in DEA formulation.

We build the DEA analysis on aggregated measures. On Fig. 9.4, all DMUs and the efficient frontier are visualised.

9.4 Conclusions

PAR (Productivity Analysis with R) is implemented in R statistical software version 2.8.1 using the DEA, FEAR and CCA packages and routines developed by us (see R Development Core Team, 2007). PAR is a very flexible, extensible software based on CCA and DEA models, implemented as CCA and FEAR packages in R. The cost of this flexibility is that the user must type commands at a command-line prompt.

In PAR methodology, CCA provides an aggregation of both input and output units and then DEA provides efficient units. The aggregation can cause significant additional bias in a DMU's technical efficiency scores. The effects of the input aggregation on efficiency indicators have been investigated. This study used data from Terceira Island. Azorean government can apply our approach to other islands and to find "the best practice" of Azorean agricultural system.

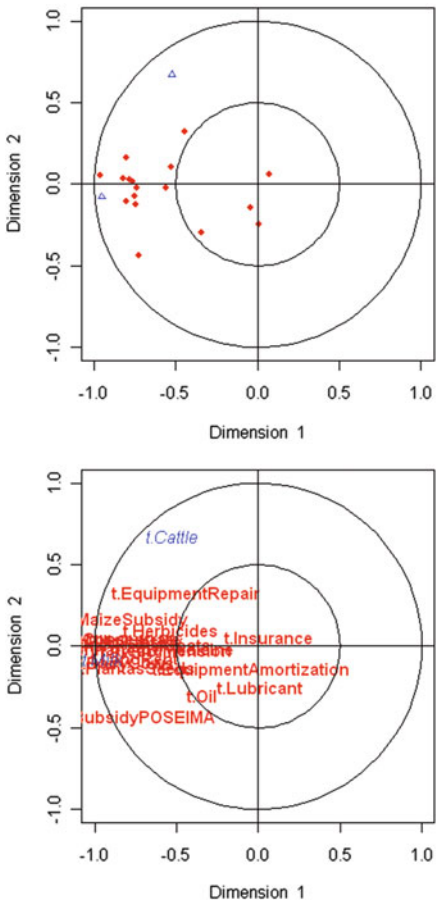

Fig. 9.3 Input and output variables plotted on the first two canonical variates

In spite of the good results achieved, it is important to recognise the major limitations and possible problems in conducting a DEA:

- Measurement error and other noise may influence the shape and position of the frontier.
- Outliers may influence the results. Because of it, we always start with outlier detection.
- The exclusion of an important input or output can result in biased results. Because of it, a variable aggregation method is proposed by PAR.
- The efficiency scores obtained are only relative to the best firms in the sample. The inclusion of extra firms (e.g. from overseas) may reduce efficiency scores.
- Be careful when comparing the mean efficiency scores from two studies. They say nothing about the efficiency of one sample relative to the other.
- The addition of an extra firm in a DEA analysis cannot result in an increase in the technical efficiency scores of the existing firms.

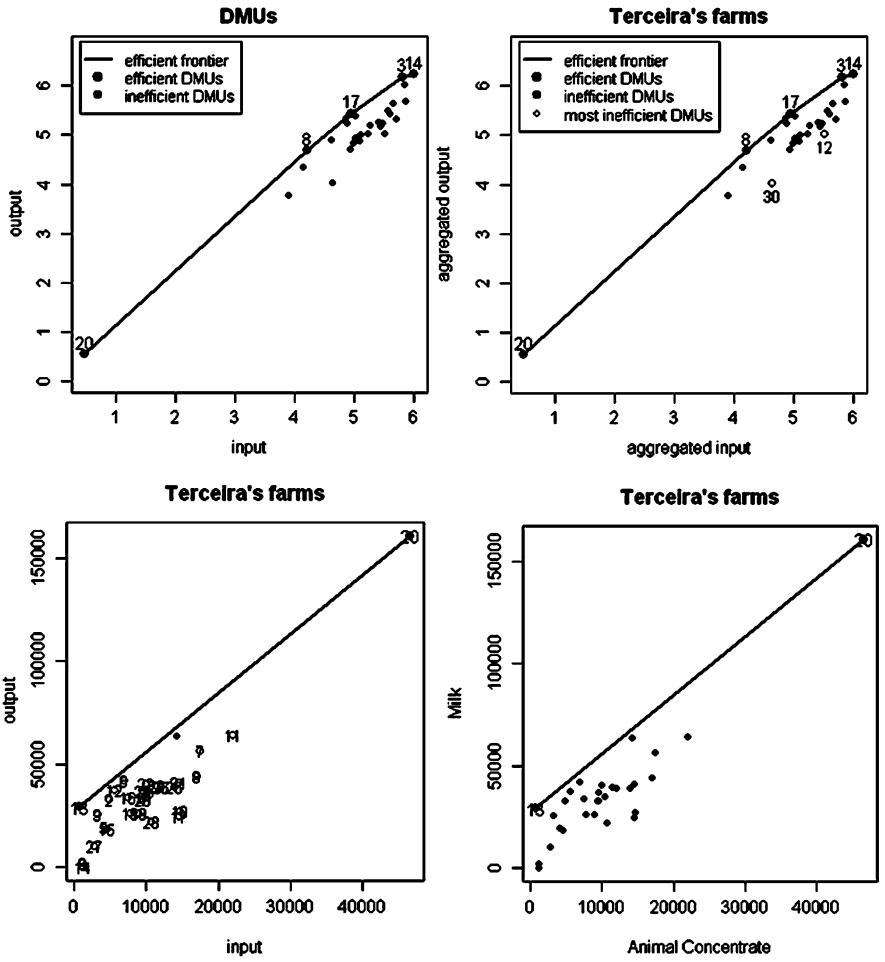

Fig. 9.4 Several examples with and without aggregation using BCC model (first two) and CCR model

- The addition of an extra input or output in a DEA model cannot result in a reduction in the technical efficiency scores.
- With few observations and many inputs and/or outputs, many of the firms will appear on the DEA frontier. If an investigator wishes to make an industry look good, he could reduce the sample size and increase the number of inputs and outputs in order to increase the technical efficiency scores. Because of it, a variable selection method is proposed by PAR.
- Treating inputs and outputs as homogeneous commodities when they are heterogeneous may bias results.

In future work, we are going to use PAR with both real and simulated data in order to find out a compromise between environment, agriculture and tourism and to investigate the potential impact of agricultural tourism on the farms' efficiency.

Acknowledgments This work has been partially supported by Regional Division for Science and Technology of Azores government through the project M.2.1.2/l/009/2008, "Productivity Analysis of Azorean Cattle-Breeding Farms with R Statistical Software".

References

Chilingerian JA (1995) Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. Eur J Oper Res 80:548–574

Cooper WW, Seiford LM, Tone K (2007) Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software, 2nd edn. Springer, New York

Diaz-Martinez Z, Fernandez-Menendez J (2008) DEA: data envelopment analysis. R package version 0.1–2

Jenkins L, Anderson M (2003) A multivariate statistical approach to reducing the number of variables in data envelopment analysis. Eur J Oper Res 147:51–61

Lewin AY, Morey RC, Cook TJ (1982) Evaluating the administrative efficiency of courts. Omega 10(4):401–411

Mendes A, Noncheva V, Silva E (2009) Decision support for enhanced productivity with R software: an Azorean farms case study. Thirty eighth annual meeting of WDSI, Hawaii, 7–11 Apr 2009

Norman M, Stoker B (1991) Data envelopment analysis: the assessment of performance. Wiley, Chichester

Nunamaker TR (1985) Using data envelopment analysis to measure the efficiency of non-profit organizations: a critical evaluation. Manag Decis Econ 6(1):50–58

Salinas-Jimenez J, Smith P (1996) Data envelopment analysis applied to quality in primary health care. Ann Oper Res 67:141–161

Sigala M, Airey D, Jones P, Lockwood A (2004) ICT paradox lost? A stepwise DEA methodology to evaluate technology investments in tourism settings. J Travel Res 43:180–192

Simar L, Wilson PW (2000) A general methodology for bootstrapping in non-parametric frontier models. J Appl Stat 27:779–802

Wilson PW (1993) Detecting outliers in deterministic nonparametric frontier models with multiple outputs. J Bus Econ Stat 11:319–323

Wilson PW (2008) FEAR: a software package for frontier efficiency analysis with R. Socio-Econ Plan Sci 42(4):247–254