Chapter 10
Sustainable Tourism and Agriculture Multifunctionality by PAR: A Variable Selection Approach

Armando B. Mendes, Veska Noncheva, and Emiliana Silva

Abstract Data Envelopment Analysis (DEA) is a popular non-parametric method used to measure efficiency. It uses linear programming to identify points on a convex hull defined by the inputs and outputs of the most efficient Decision Making Units (DMUs). Two critical elements account for the strength of the DEA approach: (1) no a priori structure is placed on the production process of the firm, and (2) the models can yield a measure of efficiency even with a very small number of data points. The first point is particularly important because the measure of efficiency is based upon the best practice of the DMUs at any of the levels of output observed.

Data envelopment analysis measures efficiency and is very sensitive to the choice of variables for two reasons: the number of efficient DMUs is directly related to the number of variables, and the selection of the variables greatly affects the measure of efficiency when the number of DMUs is few and/or when the number of explanatory variables needed to compute the measure of efficiency is too large. Our approach advises which variables should be included in a DEA model. Hence, a variable selection method is presented for the deterministic DEA approach. First, a definition of different measures of efficiency and the various DEA models used to measure efficiency is provided, and then a variable selection method is proposed. The Azorean agricultural system is used as an example to illustrate the method.

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_10,

Keywords Data envelopment analysis • Productivity analysis with R • Canonical correlation analysis • Variable selection

10.1 Introduction

Tourism is increasing in Azores islands, although in different proportion per island (S. Miguel 47%, Terceira 24.4% and Faial 12.4%). The main argument for marketing has been the nature and its conservation. The green islands can only stay green if it is possible to have a sustainable compromise between environment, agriculture and tourism.

From the SREA (2007a) characterisation of Azorean tourism, tourists were mainly elder, settled and experienced (around 45–54 years old); they come mainly from Portugal mainland, Nordic countries (Denmark, Norwegian, Sweden, Iceland) and Diaspora countries (United States of America and Canada). The most part of the tourists have higher education and a professional activity. They choose the "Azores destiny" mainly to "relax", "business" or "to visit family and friends", and they are attracted by "landscape", "nature" and "exotism" of the islands. The tourists pay, in average, per trip about 1,193€ and they stay in Azores about 9 days. The establishments preferred are "hotel" and "family and friends houses".

In 2007, the Azores had about 82 establishments (distributed for the nine islands) for agricultural tourism: 54.9% were country houses, 23.2% rural tourism, 17% lodging tourism, 3.7% agrotourism and 1.2% village tourism (SREA 2007b).

For increasing the income of agricultural enterprises, European Union had developed the concept of multifunctionality understood as "a characteristics of an activity which produces multiple and interconnects results and effects" (OECD 2001). The functions of agricultural multifunctionality are various such as agricultural, ecological, cohesion, recreational, educational, cultural and residential. In this case, the rural tourism presents as an alternative of farm income (Rodriguez et al. 2004).

The data available shows that Azores have potentiality to this kind of rural establishments. How can it affect the efficiency of dairy farms, the most representative type in Azorean agriculture? If less extensive grazing system were compensated by the increase of country houses, could this services maintain the same income and raise the efficiency? How much must an agricultural unit receive, from tourism income, to compensate the loss of income by extensive grazing system, without becoming inefficient?

Efficiency was initially measured in Azores farms by Silva et al. (2004). They had measured the Azores dairy farms' technical efficiency by applying a non-parametric efficiency analysis to a panel data of 122 dairy farms from the Azores, Portugal, for 1996. The analysis used DEA with constant and variable returns to scales models, with an input-oriented model approach. Two outputs (milk production and subsidies) and three inputs (agricultural area, number of dairy cows and variable and fixed cost) were considered relevant. The results suggest that the average technical efficiency is very low (66.4%) compared with published research data, and only a

few (7%) dairy farms were found to be efficient. In fact, the Azores dairy farms must increase their technical efficiency, given that they operate above their resource capacity. The lower efficiency showed that it is possible to produce the same amount of milk while saving approximately 33.6% of resources (or inputs).

The small dimensions (less than 25 ha per farm) may explain this low efficiency in the Azores. The Azorean farms are smaller than farms in New Zealand, Canada or Australia (Jaforullah and Whiteman 1999; Fraser and Cordina 1999; and Cloutier and Rowley 1993). The last researches suggest bigger farms are more efficient. The inefficiency in the Azorean dairy farms seems to be influenced by the great amount of fixed costs spent on agricultural equipment and animal feeding with concentrates.

In 82 milk dairy farms, Marote and Silva (2002) measured the efficiency in 3 years – 1997, 1998 and 1999 – using DEA. About 63.4, 62.2 and 70.7% of farms were efficient in the 1997, 1998 and 1999, respectively. The technical efficiency of Variable Returns to Scales (VRS) was 0.957, 0.951 and 0.960 in the 3 years mentioned. In this period, the efficiency was similar, but there was an increase of the farms' efficiency.

Later, Marote and Silva (2011) analysed the efficiency of 82 farms from 1997 to 1999 in Terceira Island (Azores archipelago) farms, using DEA. They used two models: model I considered two outputs – milk production and subsidies – and nine inputs, including dimension, animals and other variables and fixed costs. Model II considered one output – milk production – and the same nine inputs of model I.

The efficiency of farms does not improve with subsidies. This conclusion was observed by comparing efficiency using or not using subsidies, probably because the farms balanced the lowest subsidy amount with bigger milk production. This work showed that although the subsidies were very important contributors to the farms' income, their influence in efficiency was very small, which means that the efficiency and the number of farms efficient did not increase very much. Probably, the farms will have a greater efficiency if they rationalise the use of feeding and equipment costs.

Comparing this study with others in same conditions, the efficiency measured in Azores was bigger than in other regions, in part caused by the greater number of inputs. As has been shown by Suhariyanto (1999), more use of variables increases the efficiency value.

Silva and Santos (2007) measured the efficiency of 184 farms of Azores in 2002 using a different system production (milk, meat and mix: milk and meat). They used DEA and the results showed that the technical efficiency at constant returns to scales (CRS) was 63.2%, a variable returns to scales (VRS) about 71.4% and scale (SCA) 89.2% in milk production system. In the meat production system, the efficiency was greater than the milk system, a constant returns to scales (CRS – 69.4%; VRS – 82.9%), and smaller in scale efficiency (SCA – 84.2%). In the mix system production, apparently the most efficient system, the values were the biggest (CRS, 89%; VRS, 99.2%; and SCA, 89.8%) of the three systems. The number of the efficient farms was 9.8% in milk system, 11.1% in the meat system and 46.7% in the mix system.

Using a parametric approach, Venâncio and Silva (2004) measured the efficiency by a Stochastic Frontier Production (SFP) for three groups of farms, using Frontier software. The efficiency in the Faial Island (Azores archipelago) farms was higher than 80%, in the three clusters of farms (82, 93.2 and 85.1% for clusters A, B and C). The variables which contributed for inefficiency were subsidies and equipment costs. The most efficient farms were those with land rent, animal sales and bigger farms, such as those Hallam and Machado (1996) observed for the Portuguese case.

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 technical efficiency. The practice has been to select the variables by simply choosing the ones that make economic sense. The criteria for the choice of which explanatory variables (inputs or outputs) to include in a DEA model are rarely made explicit.

In this text the selection of the variables that capture most of the relationship between the inputs and outputs is explored for sustainable tourism and agriculture multifunctionality efficiency measures. Because of it, we are interested in the relationship between both the input and output sets of variables, and Canonical Correlation Analysis (CCA) would be the appropriate method of analysis. CCA is a multidimensional exploratory statistical method. More precisely, at first we would like to investigate the following questions:

1. To what extent can the set of two or more output variables be "explained" by the set of two or more input variables?
2. What contribution does a single input or output variable make to the explanatory power of the set of variables to which the variable belongs?

This chapter is focused on measuring efficiency when the number of DMUs is few and when the number of explanatory variables needed to compute the measure of efficiency is too large. Hence, a statistical approach to variable selection for the deterministic DEA models is presented.

10.2 The Efficiency Approach: Data Envelopment Analysis (DEA)

Two approaches are commonly used to measure efficiency: the parametric approach, which relies on statistical techniques to estimate the parameters of a production function, and the non-parametric approach, which compares the observed inputs and outputs of each firm with that of the most performing firms in the information set. The parametric approach has been subject to persistent criticism, centred on two points: the assumption that the production function has the same functional form for all the firms and the fact that econometric estimation of efficiency can produce biased and inconsistent parameter estimates (since an econometric measure of efficiency reflects the average performance and not the best performance).

Data Envelopment Analysis (DEA) is now the most popular method used to measure efficiency. DEA is a non-parametric method, which does not assume any specific production function. Instead, it uses linear programming to identify points on a convex hull defined by the inputs and outputs of the most efficient firms (DMU). Two critical elements account for the strength of the DEA approach: (1) no a priori structure is placed on the production process of the firm, and (2) the models can yield a measure of efficiency even with a very small number of data points. The first point is particularly important because the measure of efficiency is based upon the best practice of the DMUs at any of the levels of output observed.

For a given set of input and output variables, DEA produces a single comprehensive measure of performance called efficiency score. The CCR model (Charnes et al. 1978) formally introduced the linear programming to measure technical efficiency with the assumption of constant returns to scales. In the CCR model, DMUs adjust either their use of inputs or their outputs to reach the production frontier. The BCC model (Banker et al. 1984a, b) removed the assumption of constant returns to scales, and Charnes and Cooper (1985) proposed the additive DEA model, where both inputs and outputs can be adjusted simultaneously. All models use the distance to one of the facets of the production or cost frontier to generate an efficiency index.

This technique has useful applications in many evaluation contexts. The research presented in Chiang et al. (2004) is aimed at measuring hotel performance of International Tourist Hotels (ITHs) in Taiwan by DEA.

DEA can also be used for destination satisfaction management. The study by Sungsoo (2007), DEA Application for the Tourist Satisfaction Management, showed an application of DEA to a tourist destination, Jeju Island, suggesting that DEA was a useful tool to produce important information in managing destination for tourist satisfaction.

The aim for tourism organisations and businesses was to provide more efficient websites in order to gain competitive advantage. The study by Bauernfeind and Mitsche (2008) provides an example of how DEA can be used to assess the website's efficiency of tourism organisations.

The study by Marianna et al. (2004) proposed a way of assessing ICT (the information and communication technologies) productivity in the tourism industry using DEA. The methodology was applied in a data set from the three-star hotel sector in the United Kingdom.

In Gimenez-Garcia et al.'s (2007) study, a three-step data envelopment analysis model was used to reallocate resources in an organisational network. First, the model identified the excess resources of inefficient units and then reallocated these resources and set the output-oriented production goals for efficient units. Finally, the model recalculates improvement targets for the inefficient units based on the revised remaining resources. The procedure was applied to the analysis of 54 restaurant locations belonging to a Spanish fast-food chain. The results showed that originally efficient restaurants can improve their output by an average of 4.20% after a reallocation of inputs and that this reallocation is beneficial for the entire restaurant chain.

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:

1. DEA can handle multiple input and multiple output models.
2. DMUs are directly compared against a peer or combination of peers.
3. 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.
4. Do not need a functional form.

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:

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

Variable selection in DEA is problematic. The estimated efficiency for any DMU depends on the number of 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 DEA is applied, 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.

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.

In particular, a few researchers, such as Valdmanis (1992) and Hughes and Yaisawarng (2004), discussed the influence of variable selection on DEA results. They calculated efficiency scores by using alternative sets of variables and analysed the sensitivity of DEA efficiency scores. It is now recognised that improper variable selection often results in biased DEA evaluation results. Therefore, the appropriate variable selection is crucial for the successful application of the DEA technique.

The choice of the variable set in DEA is an empirical issue. Inclusion of many variables is not a viable option in DEA. As the number of variables in the DEA

model increases, more and more production units become efficient. On the other hand, when relevant variables are omitted, DEA underestimates efficiency, and the effect of this is more severe than when irrelevant variables are included in the DEA model. Lack of a standard-structured approach to variable selection in DEA makes the task of variable selection even more difficult.

Berger and Humphrey (1997) highlighted the difficulty of variable selection when appraising bank performance using DEA. There was no "perfect approach" on the explicit definition and measurement of the banks' input and output. Further, in choosing the variables, there were some restrictions on the type of variables since there is a need for comparable data to minimise possible bias arising from different accounting practices even among the banks that are bounded by federal bank guidelines. Indian banks were no exception.

In their paper "A Statistical Test for Nested Radial DEA Models", Pastor and Ruiz (2002) focused on analysing the marginal role of a given variable, called candidate, with respect to the efficiency measured by means of a DEA model. First, they have defined a new efficiency contribution measure (ECM), which finally compares the efficiency scores of the two radial DEA models differing in the candidate. This can be either one input or one output. Then, based on ECM, they have also approached the problem from a statistical point of view. They have developed a statistical test that allows us to evaluate the significance of the observed efficiency contribution of the candidate. Eventually, solving this test may provide some useful insights in order to decide the incorporation or the deletion of a variable into/from a given DEA model, on the basis of the information supplied by the data. Two procedures for progressive selection of variables were designed by sequentially applying the test: a forward selection and a backward elimination. These can be very helpful in the initial selection of variables when building a radial DEA model.

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 had 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 (Nunamaker 1985). In his analysis of DEA modelling, Nunamaker found that for selected DMUs, the addition of a highly correlated variable may substantially alter the DEA efficiency scores. He concluded that because a variable was redundant within a regression model did not mean that it was redundant within a DEA model. The existence of high correlation among variables did not necessarily mean that one of the variables could be excluded without changing the subsequent DEA results. Therefore, it would be unwise to rely strictly on regression and correlation analysis as a means of reducing the number of variables. At best, these quantitative techniques could assist in variable reduction. In a similar vein, Golany and Roll (1989) claim that one-at-a-time regression tests on the inputs and outputs should not be regarded as reliable rules for eliminating variables but rather as indicators for a need to examine some of the variables more closely.

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 was to apply regression and correlation analysis (Lewin et al. 1982). This approach purports those variables which were highly correlated with existing model variables. They are merely redundant and should be omitted from further analysis. Therefore, a parsimonious model typically showed generally low correlations among the input and output variables, respectively, Chilingerian (1995) and Salinas-Jimenez and Smith (1996).

One formal procedure is using a "stepwise" approach to variable selection that estimates the change in the efficiencies as variables are added or dropped from the analysis. This method is intended to produce DEA models that include only those variables with the largest impact on the DEA results. Examples showed that stepwise DEA modelling could be used on larger, realistic problems. While a stepwise procedure can inform for the effect of adding and removing variables in a DEA study, the determination of the "best" model to represent any given situation must rely on managerial judgement and knowledge of the operations of the actual situation being represented.

The authors Norman and Stoker (1991) proposed a method of adding variables to the DEA model one at a time. They started with a simple model involving one single output and one single input. Efficiencies for all the DMUs were then calculated. They claimed that high statistical correlation was an indicator that a particular variable influenced performance. A new variable was then added to the DEA model based on the correlation values and incorporated into the measure of efficiency. The process was repeated until no further influential variables remained. They did note 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).

Färe et al. (1988) consider that the basic information provided by the estimated frontier must remain unaffected by a forward selection or a backward elimination of inputs and outputs. They may be helpful when building a radial DEA model to assess efficiency.

Forward procedure is to be used when the analyst starts with a basic model consisting of the set of available variables he/she considers as essential to evaluate efficiency, and there also exists another set of variables that are thought of as possibly relevant to that end. The variable with the largest value of T statistics, if it is statistically significant, enters the model. The algorithm continues until either all variables are in the model or when at a given step the variable with the largest value of T is not statistically significant. The forward algorithm described above implicitly embodies the prior knowledge and experience of the analyst, as it requires an initial selection of the most relevant variables to define the model to start with.

Backward procedure is to be used when the analyst wonders if the specification set of a given DEA model used to evaluate the efficiency can be simplified

by eliminating some of the existing variables without significantly affecting the efficiency scores.

In general, it is not recommended that these kinds of automated procedures be used blindly to identify a "best" model because they can never replace professional judgement in the matter field. Nevertheless, they may complement this judgement with information provided by observed data.

In the backward approach, the goal of the method is to remove those variables that do not have significant influence on the efficiency. CCA advises which variable could be removed. The statistical test supports the decision maker to remove the variables. Several statistical tests which can be used to decide the incorporation of a variable into a DEA model have been proposed (Banker 1996). For instance, Brockett and Golany (1996) asserted that the distribution of efficiency scores is generally unknown and is difficult to describe in a low-dimensional parametric model, and they suggest the application of non-parametric statistical techniques based on rank statistics instead of the efficiency ratings themselves. They propose the use of the Mann–Whitney rank test to evaluate the statistical significance of the differences observed in efficiency within a DEA efficiency evaluation framework. See, for example, Simar (1996) for a discussion on some general aspects of the statistical analysis in DEA-type frontier models.

In general, it is possible to conclude that the process starts by selecting a small set of input and output items at the beginning and gradually enlarge the set to observe the effects of the added items. It is desirable that the number of DMUs (n) exceeds the sum of inputs (m) and outputs (s).

A heuristic formulae is

$$n \geq \max\left(ms, 3\left(m+s\right)\right) \tag{10.1}$$

10.3 Productivity Analysis with R (PAR): A Tool for Measuring Efficiency in Azores

In the PAR project, 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 intends to be self-explanatory. This makes the system appropriate to support public policies. PAR project is designed to provide a bridge from mathematical models to productivity study using R statistical software.

PAR methodology is aimed at:

- Designing a new "data-oriented" methodology for evaluating the performance of Azorean cattle-breeding farm system
- Offering a computer implementation of PAR methodology
- Locating efficiencies and inefficiencies and supporting public policy decisions

A natural measure of performance is a productivity ratio: the ratio of outputs over 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.

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.

The input-oriented DEA model aims to minimise inputs while satisfying at least the given output levels. The output-oriented DEA model attempts to maximise outputs without requiring more of any of the observed input variables. Dairy policy in Azorean islands depends on Common Agricultural Policy of the European Union and is limited by quotas, in the moment. This is the reason why output-oriented models are not used in this context.

10.4 Canonical Correlation Analysis in Variable Selection

The PAR approach applies CCA to select both input and output variables and to get final input and output sets, respectively. 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 and 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.

CCA finds two vectors that maximise the correlation between the linear combinations assuming that vectors a_1 and b_1 are normalised. The resulting variables U_1 and V_1 are called the first canonical variates and ρ_1 is referred as the first canonical correlation.

The canonical correlation is optimised such that the linear correlation between the two latent variables is maximised. Canonical correlation is used for many-to-many relationships. There may be more than one such linear correlation relating the two sets of variables, with each such correlation representing a different dimension by which the input set of variables is related to the output set. We use canonical correlation to explain the relation of the input and output sets of variables. For both input and output canonical variates, we assess how strongly it is related to measured variables in its own set and the set for the other canonical variates.

Wilks' lambda test is used to test the significance of the first canonical correlation. If $p < 0.05$, the two sets of variables are significantly associated by canonical correlation. Likelihood ratio test is a significance test of all sources (not just the first canonical correlation) of linear relationship between the two canonical variables. It is sometimes wrongly used as a test of the significance of the first or another single canonical correlation in a set of such functions.

Canonical correlation squared is the percent of variance in output set explained by input set of variables. In addition to asking how strong the relationship is between two latent variables, canonical correlation is useful in determining how many dimensions are needed to account for that relationship. Canonical correlation finds the linear combination of variables that produces the largest correlation with the second set of variables. This linear combination, or "root", is extracted and the process is repeated for the residual data, with the constraint that the second linear combination of variables must not correlate with the first one. The process is repeated until a successive linear combination is no longer significant.

Canonical correlation is a member of the Multiple General Linear Hypothesis (MLGH) family and shares many of the assumptions of multiple regression such as linearity of relationships, homoscedasticity (same level of relationship for the full range of the data), interval or near-interval data, untruncated variables, proper specification of the model, lack of high multicollinearity and multivariate normality for purposes of hypothesis testing. It also shares with factor analysis the need to impute labels for the canonical variables based on structure correlations, which function as a form of canonical factor loading; researchers may well impute different labels based on the same data.

As with factor analysis, there may be more than one canonical correlation, each representing an orthogonally separate pattern of relationships between the input and output variables. The maximum number of canonical correlations between two sets of variables is the number of variables in the smaller set.

The first canonical correlation is always the one which explains most of the relationship. The canonical correlations are interpreted in the following way: the square of the canonical correlation is the percent of variance in the canonical variate of the output set of variables explained by the canonical variate for input set. Another way to put it is to say that Rc squared is the percent of variance shared by the canonical variates along this dimension. Pooled Rc2 (pooled canonical correlation) is the sum of the squares of all the canonical correlation coefficients, representing all the orthogonal dimensions in the solution by which input and output sets of variables are related. Pooled Rc2 is used to assess the extent to which one set of variables can be predicted or explained by the other set.

The standardised canonical weights are used to assess the relative importance of an individual variable's contributions to a given canonical correlation. The canonical coefficients are the standardised weights in the linear equation of variables which creates the canonical variables. If an independent variable is totally redundant with another independent variable, its partial coefficient (canonical weight) will be zero. Nonetheless, such a variable might have a high correlation with the canonical variable (i.e. a high structure coefficient).

However, Levine (1977) argues against the procedure above on the ground that the canonical coefficients may be subject to multicollinearity, leading to incorrect judgements. Also, because of suppression, a canonical coefficient may even have a different sign compared to the correlation of the original variable with the canonical variable. Therefore, instead, Levine (1977) recommends interpreting the relations of the original variables to a canonical variable in terms of the correlations, which are called structure correlation coefficients, also known as canonical factor loadings, that is, the correlation of canonical variable scores for a given canonical variable with the standardised scores of an original input variable. The table of structure correlations is sometimes called the factor structure. In summary, the canonical weights have to do with the unique contributions of an original variable to the canonical variable, whereas the structure correlations have to do with the simple, overall correlation of the original variable with the canonical variable.

Alpert and Peterson (1972) noted that canonical weights appear more suitable for prediction, while structure coefficients may better explain underlying (although interrelated) constructs. Variables with correlations of 0.3 or above are interpreted as being part of the canonical variable, and those below are not considered part of the canonical variable.

It is well known that because the weights are partial coefficients whereas the canonical factor loadings are not, if a given variable shares variance with other independent variables entered in the linear combination of variables used to create a canonical variable, its weight is computed based on the residual variance it can explain after controlling for these variables. If an independent variable is totally redundant with another independent variable, its canonical weight will be zero. Nonetheless, such a variable might have a high correlation with the canonical variable (i.e. a high structure coefficient). In summary, the canonical weights have to do with the unique contributions of an original variable to the canonical variable, whereas the structure correlations have to do with the simple, overall correlation of the original variable with the canonical variable.

The canonical coefficients are standardised coefficients, and their magnitudes can be compared. However, Levine (1977) argues against this procedure on the ground that the canonical coefficients may be subject to multicollinearity, leading to incorrect judgements. Also, because of suppression, a canonical coefficient may even have a different sign compared to the correlation of the original variable with the canonical variable. Therefore, instead, Levine recommends interpreting the relations of the original variables to a canonical variable in terms of the correlations of the original variables with the canonical variables – that is, by structure coefficients. This is the standard approach.

The CCA assumptions are:

1. Interval level data are assumed.
2. Linearity of relationships is assumed, though there are nonlinear canonical correlation procedures like OVERALS algorithm (Gifi 1990). In the usual form of canonical correlation, however, analysis is performed on the correlation or variance-covariance matrices, which reflect linear relationships. Of course, one

can insert exponentiated or otherwise nonlinearly transformed variables into either measured variable set in canonical correlation.

3. Low multicollinearity: To the extent that the variables within the independent sets of variables are highly correlated, the canonical coefficients will be unstable. The coefficients for some variables may be misleadingly low or even negative because variance has already been explained by other variables.
4. Homoscedasticity and other assumptions of correlation are assumed. The covariates are created based on the correlation matrix, with regression-like assumptions that the degree of correlation is constant along the full range of the variables being correlated.
5. Minimal measurement error is assumed since low reliability attenuates the correlation coefficient. Canonical correlation can also be quite sensitive to missing data.
6. Unrestricted variance: If variance is truncated or restricted due, for instance, to poor sampling, this can also lead to attenuation of the correlation coefficient.
7. Similar underlying distributions are assumed: If two variables come from unlike distributions, their correlation may be well below +1 even when data pairs are matched as perfectly as they can be, while still conforming to the underlying distributions. That is, the larger the difference in the shape of the distribution of the two variables, the more the attenuation of the correlation coefficient. This assumption may well be violated when correlating an interval variable with a dichotomy or even an ordinal variable.
8. Multivariate normality is required for significance testing in canonical correlation. This assumption is violated when dichotomous, dummy and other discrete variables are used. In such situations, where significance testing is not appropriate, researchers may use a resampling method. The central limit theorem demonstrates, however, that for large samples, indices used in significance testing will be normally distributed even when the variables themselves are not normally distributed, and therefore, significance testing may be employed.
9. Non-singularity in the correlation matrix of original variables. This is the problem of perfect multicollinearity: a unique solution cannot be computed if some variables are redundant, thereby approaching perfect correlation with others in the model. A correlation matrix with redundancy is said to be singular or ill conditioned. Data sets based on survey data, in which there are a large number of questions, are more likely to have redundant items.
10. Adequate sample size must exist to reduce the chances of type II error (thinking you don't have something when you do). Stevens (1986) recommends at least 20 times as many cases as variables in the analysis in order to interpret the first canonical correlation only. For two canonical correlations, Barcikowski and Stevens (1975) recommend 40–60 times as many cases as variables.
11. No or few outliers. Outliers can substantially affect canonical correlation coefficients, particularly if sample size is not very large.

We will investigate the relationship between input and output sets of variables. More precisely, we would like to investigate the following questions:

1. To what extent can the set of two or more output variables be "explained" by the set of two or more input variables?
2. What contribution does a single input or output variable make to the explanatory power of the set of variables to which the variable belongs?

10.5 The Example of Azorean Farms' Efficiency

The Azores islands belong to the Portuguese territory with a population of about 250,000 inhabitants; most part (about 75%) of this population is in S. Miguel and Terceira islands.

The main economic activity is dairy farming. Azores also produce wine and vegetables. In smaller islands, the meat production was more important than milk as is the case for Santa Maria and Corvo, but other agricultural productions were residual. In consequence, it is no surprise that milk production was the major product in value over 70.4% in relation to all farm production, in 1997, which has been improving to 76.2% in 1998 and 79.3% in 1999 (Marote and Silva 2011).

Azorean dairy policy depends on Common Agricultural Policy (CAP) of the European Union, namely, the milk production quota (about 547 million tons in 2011), although the Regional Government of Azores can adjust some rules on the regional agricultural policy. Some examples are measures to combat plagues and diseases, support to specific cultures needed for industrial transformation as sugar beet for sugar production and subsidies for biological production.

Azorean farms are small; some Azorean statistics shows about 8 ha per farm, about the half of the European average dimension (15.8 in 2003). The production system is primarily based on grazing (about 95% of the area). There were about 15,107 farmers in Azores. They were mainly old, more than 55 years old and low educated, mainly with basic education (4 years). This characterisation of the Azorean farms was based on national agricultural institution data and previous works of the authors (Silva et al. 1996).

The current historical context is particularly complex as some major changes are likely to occur. This is the case for the increased prices of some food products in international markets and, locally, the end of milk quota system. The multiplying effect of agriculture in both a small economy and the Azorean society makes this kind of work of major interest not only to protect the income of farmers but also to keep the society in equilibrium on employment matters and reduce immigration cycles. In this context, decision makers need information and knowledge for deciding the best policies in promoting quality and best practices.

The most important cost in Azores bovine farms are concentrates (variable cost and annual amortisation, fix cost). For 2002 data, the concentrates have the biggest value (about 30%) in milk systems and the lowest value in beef farms. The annual depreciation have increased in the last years and reached about 20% of total costs.

Fertilisers and land rent follow the importance in total cost (about 10%). The fertilisers get importance in beef systems with 16.6% of total costs and balanced the less use of concentrates (7.9%). The conservation and repair of equipment and construction vary from 4.9 (milk system) to 8.4% (mix system). The use of oil and petrol has values between 5.3 and 10.8%. Because the discriminated costs have more importance in total structure costs, it is supposed that they will affect more the efficiency of farms and they must be considered as variables for defining efficiency or inefficiency in farms.

The subsidies were an important part of the profit of dairy farms, and in 2004, it was about 61.6% of all profits in average per farm. Azorean agricultural farms had five main kinds of support from EU and government:

- Support to limit a plus production
- Maintaining extensification production, lake protection, protection of the genetic variability, etc.
- Money for more ecological production (the agri-environment measures)
- Support investment for planting trees in previously cultivated areas, maintaining tree culture, and support for less production as a result of planting trees
- Early retirement scheme

We investigate all 30 animal farms from Terceira Island of Azores. The initial list of potential variables is large. Any resource used by an Azorean dairy farm is treated as an input variable. The output variables come from the performance and activity measures that result when a farm converts resources to produce products. Following (Boussofiane et al. 1991) environmental variables which add resources are treated as inputs in our DEA models whereas those that require resources are treated as outputs. Applying DEA procedure, we focus on the choice of data variables in addition to the methodology of DEA.

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*. After the outlier detection, one outlier identified in Terceira data was the result of a recording error and was corrected.

Canonical correlation analysis aims at highlighting correlations between input and output data sets. Two preliminary steps calculate the sample correlation coefficients and visualise the correlation matrixes. The correlation matrixes are visualised in Fig. 10.1.

Figure 10.1 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.

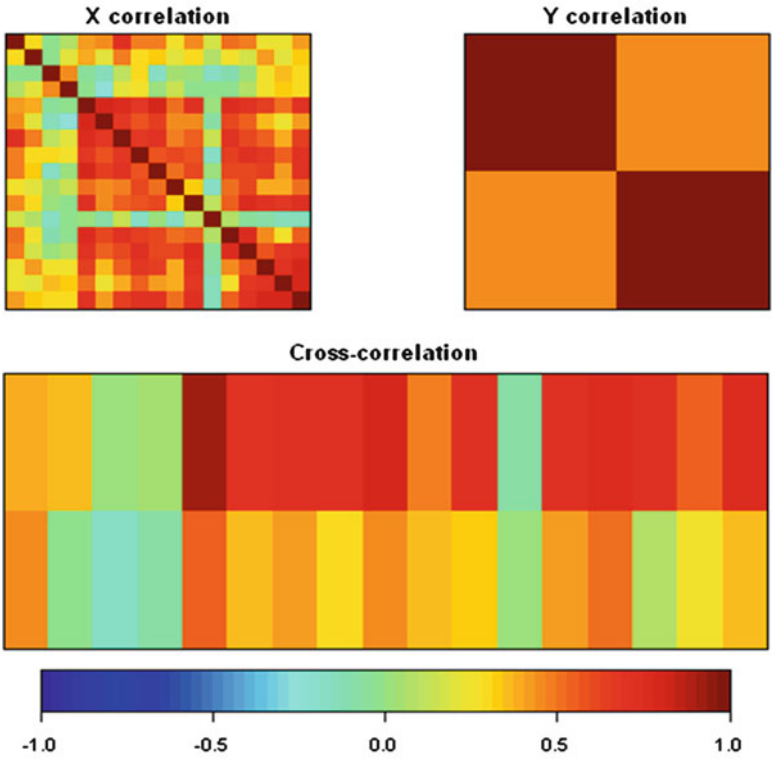

Fig. 10.1 Visualisation of sample correlation coefficients

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 were already presented in the previous chapter of this book (Noncheva et al. 2012). From these results, we can conclude that both canonical variates were predominantly associated with the following original inputs: *AnimalConcentrate*, very big value almost a unit (−0.96); *Fertilizers* (−0.82); and with the original output variable *Milk*. These three variables are very strongly correlated, meaning that they contain the strongest dependence relation; they should be the major discriminant factors between farms. This way we selected the following two input variables *AnimalConcentrate* and *Fertilizers* and one output variable *Milk*. Note that, because the outputs have negative relations with the variates, negative values of the inputs should be directly related with the inputs and positive values are inversely related. There is very little negative correlation in the inputs and all are very weak, meaning that almost all the inputs contribute positively to the outputs.

The subsidies are very important in the farm income's output, especially maize and milk subsidies, respectively, −0.80 and −0.79. This result is also corroborated by statistical data as the Azorean SREA (2007b), where we can find that about 25% of the incomes of farmers were subsidies; this is a very big slice of the income. Also, if the subsidies were eliminated, the farm profit per month would be less than 400€ (inferior to minimum salary in Portugal). This shows the great importance of the subsidies in the sustainability of Azorean farms.

The number of dairy cows is more important than the farm dimension, respectively, −0.81 and −0.56, as it was expected because the number of cows should be more directly related with the milk production than the number of hectares, in spite of the effort towards extensive explorations.

In the previous chapter, "Azorean agriculture efficiency by PAR", we explore the concept of using CCA for aggregation of inputs and outputs. Here we use the same data to variable selection. Very similar results can be obtained.

10.6 Final Remarks

DEA models are used by PAR methodology to measure efficiency in production of Azoreans farms. DEA models are useful in situations in which multiple outputs are produced from a vector of inputs and no reliable price information exists that would allow estimation of stochastic frontier cost functions (Lovell 1993).

The "Productivity Analysis with R" (PAR) framework establishes a user-friendly data envelopment analysis environment with special emphasis on variable selection and aggregation and summarisation and interpretation of the results. The starting point is the following R packages: DEA (Diaz-Martinez and Fernandez-Menendez 2008) and FEAR (Wilson 2005). The DEA package performs some models of data envelopment analysis presented in Cooper et al. (2007). FEAR is a software package for computing non-parametric efficiency estimates and testing hypotheses in frontier models. FEAR implements the bootstrap methods described in Simar and Wilson (2008).

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

It was applied, the PAR framework, to estimate the efficiency of the agricultural system in Azores (Noncheva et al. 2009). It is possible to rank observations (Azorean farms) in terms of their dissimilarity to other observations in the data (other Azorean farms). This makes PAR appropriate to support public policies in agriculture sector in Azores.

It is offered as a formal procedure for our intuitively sound approach to variable selection. This approach looks at the changes in the DMU's efficiencies when variables are added and removed from the DEA models in order to determine whether these changes are statistically significant. The procedure for variable selection has been developed by sequentially applying statistical and DEA techniques. This procedure is intended to produce DEA models that include only those variables that contribute to the closer input/output relations and have largest impact on the DEA results.

While this formal procedure can inform for the effect of adding and removing variables in a DEA study, the determination of the "best" model to represent any given situation must rely on managerial judgement and knowledge of the operations of the actual situation being represented.

It starts by selecting an initial model, involving all input and output variables. Next, the efficiency estimates for the initial model were compared to those for a new model in which some variables were subtracted. Efficiencies are calculated for each DMU under both the initial and reduced model. A statistical test was performed to determine whether the subtracting of some variables would significantly decrease the efficiency estimates. This procedure can be repeated until we receive a parsimonious model, using as many variables as needed but as few as possible.

PAR is used 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 Direcção Regional da Ciência e Tecnologia of Azores Government through the project M.2.1.2/I/009/2008.

References

Alpert M, Peterson R (1972) On the interpretation of canonical analysis. J Mark Res 30:29–50

Banker R (1996) Hypothesis tests using data envelopment analysis. J Product Anal 7(2–3):139–159

Banker R, Charnes R, Cooper W (1984a) Some models for estimating technical and scale inefficiencies in data envelopment analysis. Manag Sci 30:1078–1092

Banker R, Charnes R, Cooper WW (1984b) Equivalence and interpretation of alternative methods for determining returns to scales in data envelopment analysis. Eur J Oper Res 89:473–481

Barcikowski R, Stevens J (1975) A Monte Carlo study of the stability of canonical correlations, canonical weights, and canonical variate-variable correlations. Multivar Behav Res 10:353–364

Bauernfeind U, Mitsche N (2008) The application of the data envelopment analysis for tourism website evaluation. Inf Technol Tour 10(13):245–257

Berger A, Humphrey D (1997) Efficiency of financial institutions: international survey and directions for future research. Eur J Oper Res 98:175–212

Boussofiane A, Dyson RG, Thanassoulis E (1991) Applied data envelopment analysis. Eur J Oper Res 52:1–15

Brockett PL, Golany B (1996) Using rank statistics for determining programmatic efficiency differences in data envelopment analysis. Manag Sci 42(3):466–472

Charnes A, Cooper WW (1985) Foundations of data envelopment analysis for Pareto-Koopmans efficient empirical production functions. J Econ 30(1/2):91–107

Charnes A, Cooper W, Rhodes E (1978) Measuring the efficiency of decision-making units. Eur J Oper Res 2:429–444
Chiang WE, Tsai M-H, Wang LS-M (2004) A DEA evaluation of Taipei hotels. Annal Tour Resear 31(3):712–715
Chilingerian J (1995) Evaluating physician efficiency in hospitals: a multivariate analysis of best practices. Eur J Oper Res 80:548–574
Cloutier L, Rowley R (1993) Relative technical efficiency: data envelopment analysis and Quebec's dairy farms. Can J Agric Econ 41:169–176
Cooper W, Seiford L, 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) The DEA package, Version 0.1–2, Retrived from http://cran.r-project.org/web/packages/DEA/DEA.pdf
Färe R, Grosskopf S, Lovell C (1988) Scale elasticity and scale efficiency. J Inst Theor Econ 144:721–729
Fraser I, Cordina D (1999) An application of data envelopment analysis to irrigated dairy farms in northern Victoria, Australia. Agric Syst 59:267–282
Gifi A (1990) Nonlinear multivariate analysis. Wiley, Chichester
Gimenez-Garcia VM, Martínez-Parra JL, Frank P (2007) Improving resource utilization in multi-unit networked organizations: the case of a Spanish restaurant chain. Tour Manag 28:262–270
Golany B, Roll Y (1989) An application procedure for DEA. Technion-Israel Institute of Technology, Israel
Hallam D, Machado F (1996) Efficiency analysis with panel data – a study of Portuguese dairy farms. Eur Rev Agric Resou Econ 23(1):79–93
Hughes A, Yaisawarng S (2004) Sensitivity and dimensionality tests of DEA efficiency scores. Eur J Oper Res 154:419–422
Jaforullah M, Whiteman J (1999) Scale efficiency in the New Zealand dairy industry: a non-parametric approach. Aust J Agric Resour Econ 43(4):523–541
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
Levine M (1977) Canonical analysis and factor comparison, vol 6, Quantitative applications in the social sciences series. Sage Publications, Thousand Oaks
Lewin A, Morey R, Cook T (1982) Evaluating the administrative efficiency of courts. Omega 10(4):401–411
Lovell C (1993) Production frontier and productive efficiency. In: Fried HO, Lovell CAK, Schmidt SS (eds) The measurement of productive efficiency-techniques and applications. Oxford University Press, Oxford, pp 3–67
Marianna S, David A, Peter J, Andrew L (2004) ICT paradox lost? A stepwise DEA methodology to evaluate technology investments in tourism settings. J Trav Resear 43(2):180–192
Marote E, Silva E (2002) Análise Dinâmica da Eficiência das Explorações Leiteiras da Ilha Terceira. XII Congresso de Zootecnia. November
Marote E, Silva E (2011) Importância dos Subsídios na Eficiência das Explorações Leiteiras da Terceira. Revista de Ciências Agrárias, pp 161–170
Noncheva V, Mendes A, Silva E (2009) An approach to variable aggregation in efficiency analysis. In: Classification, forecasting, data mining, international book series information science & computing. Suppl Int J Inf Technol Knowl 3(8):97–104
Noncheva V, Mendes A, Silva E (2012) Azorean agriculture efficiency by PAR. In: Mendes A, Silva E, Santos J (eds) Efficiency measures in the agricultural sector, with applications. Springer, Dordrecht, pp 53–72
Norman M, Stoker B (1991) Data envelopment analysis: the assessment of performance. Wiley, Chichester
Nunamaker T (1985) Using data envelopment analysis to measure the efficiency of non-profit organizations: a critical evaluation. Manag Decis Econ 6(1):50–58

OECD – Organisation for Economic Co-Operation and Development (2001) Environmental indicators for agriculture methods and results. Executive summary. Retrived from http://www.oecd.org/greengrowth/sustainableagriculture/1916629.pdf.

Pastor JT, Ruiz I-S (2002) A statistical test for nested radial DEA models. Oper Res 50(4):728–735

Rodriguez M, Gómez E, Lorente J (2004) Rural multifunctionality in Europe. The concepts and policies. 90th AEEA Seminar

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

Silva E, Santos C (2007) Eficiência nos Sistemas de Produção Pecuária nos Açores. APDEA Congress, Vila Real

Silva E, Arzubi A, Berbel J (1996) An application of data envelopment analysis (DEA) in Azores dairy farms. New Medit 3:39–43

Silva E, Arzubi A, Berbel J (2004) An application of data envelopment analysis (DEA) in Azores dairy farms. New Medit 3:39–43

Simar L (1996) Aspects of statistical analyses in DEA-type frontier models. J Product Anal 7:177–186

Simar L, Wilson P (2008) Statistical interference in nonparametric frontier models: recent developments and perspectives. In: Fried H, Lovell CAK, Schmidt S (eds) The measurement of productive efficiency and productivity change. Oxford University Press, New York

SREA (2007a) Estudo sobre os Turistas que visitam os Açores. 2005 – 2006. Região Autónoma dos Açores/ed. Serviço Regional de Estatística dos Açores

SREA (2007b) Anuário Estatístico dos Açores, 2007. Região Autónoma dos Açores/ed. Serviço Regional de Estatística dos Açores

Stevens J (1986) Applied multivariate statistics for the social sciences. Erlbaum, Hillsdale

Suhariyanto K (1999) Productivity growth efficiency and technical changes in Asian agriculture: a Malmquist index analysis. PhD thesis, University of Reading

Sungsoo P (2007) DEA application for the tourist satisfaction management. Tour Anal 12:201–211

Valdmanis V (1992) Sensitivity analysis for DEA model: an empirical example using public vs. NEP hospitals. J Public Econ 48:185–205

Venâncio F, Silva E (2004) A Eficiência de Exploração Agro-pecuárias dos Açores: uma abordagem paramétrica. XIV Jornadas Luso Espanholas de Gestão Científica

Wilson PW (2005) FEAR 1.0: a software package for frontier efficiency analysis with R, Retrieved from http://business.clemson.edu/Economic/faculty/wilson/courses/bcn/papers/fear.pdf