CATEGORICAL VARIABLES IN DEA[1]

by

Finn R. Førsund

Department of Economics, University of Oslo, Norway

and

Visiting Fellow ICER, Turin, Italy

Email: f.r.forsund@econ.uio.no

March 2001

Abstract: If a DEA model has a mix of categorical and continuous variables a standard LP formulation can still be used by entering all combinations of categorical and continuous variables as different types of inputs and/or outputs. Most units will then not have positive levels of all variables. The implications for selection of peers are investigated. Peers can have the same or fewer types of inputs than the unit under investigation, but either fewer or more types of outputs. There is a basic asymmetry between number of positive inputs and outputs of the peer units due to more of inputs reducing efficiency while more of outputs improving efficiency. The special cases of imposing a hierarchical structure on the categorical variables dealt with in the literature can easily be incorporated.

Key words: Categorical variable, DEA, efficiency, linear programming, peer.

JEL classification: C6, D2

[1] The paper is written while visiting fellow at International Centre for Economic Research (ICER), Turin, January – March 2001. It is part of the project "Cheaper and better?" at the Frisch Centre, financed by the Norwegian Research Council**.**

1. Introduction

In efficiency analysis of production units, called *DMU*s (decision-making units in the operations research literature), resources consumed and outputs produced are usually assumed to be continuous variables. However, in practical applications some variables may be *categorical*. A categorical variable is a variable that takes on only a finite number of values. It is not unusual in DEA applications, especially for DMUs where outputs are not sold on markets, that variables are categorical, i.e. there are inputs and/or outputs of certain types, e.g. type of education of labour, type of court cases being completed, etc., which are distinct and cannot be represented by continuous variables.

DEA models with categorical variables are treated for the first time (to our knowledge) in Banker and Morey (1986), and the approach improved by Nakamura (1988) and also followed up in Rousseau and Semple (1993). Charnes et al. (1994) provide a further development, which is used in e.g. Puig-Junoy (1998), and also presented in Cooper et al. (2000). However, the programming models developed in the first two papers are of the mixed integer type. Moreover, all papers mentioned are concerned only with ordered categorical variables (e.g. of the type "low", "medium", "high"). Our purpose is to show how to adapt a standard LP programme formulation of the DEA model, as done in Charnes et al. (1994), but not restricted to hierarchically ordered variable types only. The crucial assumption needed is that there is at least one continuous input variable and at least one continuous output variable.

There may be important applications of DEA analyses where imposing a hierarchical structure would not be natural. In an efficiency analysis of the municipal nursing and home care sector Erlandsen and Førsund use number of clients in a limited set of age groups as categorical output variables, as well as whether nursing homes have single rooms or not (see Erlandsen and Førsund, 2001). In a study of the efficiency of auction houses of selling Picasso paintings, Førsund and Zanola (2001) use Picasso paintings from different periods in the painter's life as both categorical inputs and outputs. Such variables have no natural ordering.

An important output of a DEA efficiency analysis is the identification of peers for inefficient units. Due to the general feature of models with categorical variables that the DMUs often do not have full sets of positive variables, it is of interest to study the nature of peers with respect to positive variables. The situation with hierarchically ordered categorical variables would come out as a special case. The DEA models for calculating efficiency scores are set out in Section 2 and our general way of treating categorical variables is developed in Section 3. A review of the literature in view of our generalisation is provided in Section 4, and some concluding remarks offered in Section 5.

2. The DEA efficiency model

The point of departure for the calculation of efficiency measures is the piecewise linear frontier technology expressed by the following production possibility set:

$$S = \{(x, y) : y\, can\, be\, produced\, by\, x \ \} =$$
$$\{(x, y) : \sum_{j=1}^{J} \lambda_j y_{mj} \geq y_m \forall m\, ,\ x_n \geq \sum_{j=1}^{J} \lambda_j x_{nj} \ \ \forall n\, ,\ \lambda_j \geq 0\ \forall j\} \tag{1}$$

where x is the input vector and y is the output vector, and in the last expression we have introduced J points and index m for type of output and index n for type of input. The variables λ_j (j=1,..,J) are non-negative weights or intensity variables defining frontier points. Constant returns to scale is assumed[2]. Basic properties are that the production set is convex, includes all points and envelopment is done with minimum extrapolation, i.e. the fit is as "tight" as possible.

The input oriented Farrell radial efficiency measure, E_{1i}, for each DMU, i, of a set of J observations, is calculated by solving the following linear programme set up according to the definition of the measure:

[2] The nature of scale does not play any part in determining the type of the LP programme or the solutions with categorical variables.

$$\begin{aligned}
&E_{1i} = Min\ \theta_i \\
&s.t. \\
&\sum_{j=1}^{J} \lambda_j y_{mj} - y_{mi} \geq 0\ ,\ m = 1,..,M \\
&\theta_i x_{ni} - \sum_{j=1}^{J} \lambda_j x_{nj} \geq 0\ ,\ n = 1,..,N \\
&\lambda_j \geq 0\ ,\ j = 1,..,J
\end{aligned} \tag{2}$$

Each type of input is scaled down with the same factor, θ_i, until the frontier is reached according to the definition of the Farrell efficiency measure. DMUs with positive λ_j are termed "Peers". These DMUs have to be frontier units, and the linear combinations define the frontier point that is the point of comparison with the DMU_i under investigation. In the case of zero slacks on the output constraints the comparison point is found as a radial contraction of the observation for DMU_i. In case of slacks the frontier point will coincide with one of the peers.

The output oriented Farrell radial efficiency measure, E_{2i}, for each unit, i, of a set of J observations, is calculated by solving the following linear programme set up according to the definition of the measure, with the necessary change that we solve for the inverse measure $\phi_i = 1/ E_{2i}$ in order to maintain a linear programming problem:

$$\begin{aligned}
&\frac{1}{E_{2i}} = Max\ \phi_i \\
&s.t. \\
&\sum_{j=1}^{J} \lambda_j y_{mj} - \phi_i y_{mi} \geq 0\ ,\ m = 1,..,M \\
&x_{ni} - \sum_{j=1}^{J} \lambda_j x_{nj} \geq 0\ ,\ n = 1,..,N \\
&\lambda_j \geq 0\ ,\ j = 1,..,J
\end{aligned} \tag{3}$$

Each type of output is scaled up with the same factor, ϕ_i, until the frontier is reached according to the (inverse) definition of the Farrell efficiency measure. DMUs with positive λ_j (for convenience the same symbol is used as in the input-oriented case) are, again, the peers. These DMUs have to be frontier units, and the linear combinations define the frontier point that is the point of comparison with the DMU$_i$ under investigation. In the case of zero slacks on the output constraints the comparison point is found as a radial expansion of the observation for DMU$_i$. In case of slacks the frontier point will coincide with one of the peers.

It should be noted that a general assumption underlying the rationale for comparing different production units is that the inputs and outputs are indeed comparable, i.e. that they are homogeneous. If x_{ni} is labour input measured in hours then these hours must be comparable across the units. It would not be so meaningful an analysis if one unit has highly educated employees while another has unskilled ones if we believe that marginal productivity of these two types of labour are significantly different.

3. Features of the DEA solution with categorical variables

Handling of categorical variables

In the DEA model a general way of handling categorical variables may be to interpret the different attributes or states as different *types* of inputs and/or outputs, recognising the need for homogenous variables across DMUs. Let z_{kj}^{x} be a categorical characteristic k (k=1,.., K) of DMU$_j$ (j=1,..,J) regarding types of inputs, z_{pj}^{y} be a categorical characteristic p (p=1,..,P) of DMU$_j$ regarding types of outputs, and let x_{nj} be continuous input variables of type n (n=1,..,N), and y_{mj} be continuous output variables of type m (m=1,..,M) . We then have KxN different types of inputs, each continuous input variable is matched with each of the K types of inputs, and PxM different types of outputs, each continuous output variable is matched with each of the M types of outputs. Thus, the situation with a mix of categorical and continuous variables is converted to a standard DEA LP model[3].

Each DMU may typically employ fewer characteristics than the total number existing, resulting in a value of zero for the non- observed types of inputs. An extreme case would be that each DMU employs only one type of input, e.g. only labour of one

[3] Note that we may run into a dimensionality problem as to number of observations and number of variables.

category of education, and then there may be a number of combinations of having and not having certain variables. This situation imposes some restrictions on what kind of peers that will emerge. We now go on to explore these restrictions.

Input orientation

Let us look at the input restrictions in the efficiency score programme (2) above and reinterpret the number of inputs, N, as including all categorical variables converted to homogeneous types. We assume that each type of input is employed at least by one unit. The production unit under investigation is DMU$_i$. The restriction system for inputs is:

$$\begin{aligned} &\theta_i x_{1i} - \sum_{j=1}^{J} \lambda_j x_{1j} \geq 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\theta_i x_{ni} - \sum_{j=1}^{J} \lambda_j x_{nj} \geq 0 \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ &\theta_i x_{Ni} - \sum_{j=1}^{J} \lambda_j x_{Nj} \geq 0 \end{aligned} \tag{4}$$

If DMU$_i$ is not using type n input then the corresponding constraint is:

$$-\sum_{j=1}^{J} \lambda_j x_{nj} \geq 0 \tag{5}$$

Both λ_j and x_{nj} are non-negative variables. Fulfilment of the constraint then requires all the products of λ_j and x_{nj} to be zero. If x_{nj} is positive for a type j (i.e. we are considering an input type the unit under investigation does not have) then the corresponding λ_j must be zero. This implies that the peers cannot have positive x_{nj}. The peers cannot employ types of inputs the unit under investigation are not using[4].

[4] A similar result may be deduced from the analysis in Banker and Morey (1986), p.1618.

For the case that DMU$_i$ has input of type n, but that some of the peer DMUs do not have this type, we may denote the set of DMUs with this input n for Jn' and the set of DMUs without for J_n". The constraint then reads:

$$\begin{aligned} &\theta_i x_{ni} - \sum_{j \in J_n'} \lambda_j x_{nj} - \sum_{j \in J_n''} \lambda_j x_{nj} \geq 0 \\ &\Rightarrow \theta_i x_{ni} - \sum_{j \in J_n'} \lambda_j x_{nj} \geq 0 \qquad , n = 1,..,N \end{aligned} \tag{6}$$

The x$_{ni}$ are positive by assumption. Is it possible that none of the peers employ a factor of type n? Equation (6) will hold with inequality even if all the peer variables of input type n should be zero. The optimal value of the efficiency score must then be determined from other binding input constraints. If the constraint (6) is not binding it cannot influence the solution for the weights and the efficiency score. The weight is the same for all inputs and outputs of a peer unit, j. Notice that no unit can have all inputs zero and still have positive outputs by assumption on the production set (1). A peer must then at least have one type of input in common with the unit under investigation, since by Equation (5) we have that a peer cannot employ input types the unit under investigation does not use.

The efficiency score is calculated from a binding constraint of type (6):

$$\begin{aligned} &\theta_i x_{ni} - \sum_{j \in J_n'} \lambda_j x_{nj} = 0 \\ &\Rightarrow \theta_i = \frac{\sum_{j \in J_n'} \lambda_j x_{nj}}{x_{ni}} \qquad , n \in N' , \end{aligned} \tag{7}$$

where N' is the set of inputs with a binding constraint in (6). We need at least one binding constraint to calculate an efficiency score. For the case of the unit under investigation being inefficient we also need at least one other constraint in (2) to hold with equality in order to determine at least one positive weight, λ_j.

Let us similarly reinterpret the number of outputs, M, to include all categorical variables converted to homogeneous types. The constraint for an output type, m, not produced by the unit under investigation reads:

$$\sum_{j=1}^{J} \lambda_j y_{mj} \geq 0 \tag{8}$$

This constraint does not exclude peers from having positive amounts of an output m that the unit under investigation does not produce. If that should be the case the constraint (8) is not binding and thus cannot influence the solution for the weights and the efficiency score.

For the case that DMU$_i$ has output of type m, but that some of the potential peer DMUs do not have this type, we may now denote the set of DMUs with this output m for Jm' and the set of DMUs without for J$_m$". The constraint then reads:

$$\begin{aligned} &\sum_{j \in J_m'} \lambda_j y_{mj} - \sum_{j \in J_m''} \lambda_j y_{mj} - y_{mi} \geq 0 \\ &\Rightarrow \sum_{j \in J_m'} \lambda_j y_{mj} - y_{mi} \geq 0 \quad , m = 1,..,M \end{aligned} \tag{9}$$

A peer may according to (9) have fewer types of outputs than the unit under investigation. Since there is only one weight for each unit we must have at least two constraints of type (6) or (9) to hold with equality to determine a weight and the efficiency score. A peer unit must be involved in at least one binding constraint of type (6) or (9) for a positive weight to be determined. In general we have as the maximal number of non-negative solutions for the efficiency score and the weights the number of constraints that are binding. We cannot have a feasible solution with just the efficiency score positive and all weights zero. The extreme case is a unit becoming a self-evaluator where only the weight for the self-evaluator becomes positive, and equal to one and the same value for the efficiency score. All the (MxN) constraints in (2) are then binding, but only two endogenous variables have positive solutions.

Is it possible that no peers have an output of type m? If that should be the case we must have:

$$-y_{mi} \geq 0 \ , m = 1,..,M \tag{10}$$

But this is not possible by the assumption of variables being non-negative, so we can conclude that for outputs it is necessary that at least one peer is producing the same output as the unit under investigation for each of its outputs. The frontier must be fully faceted in the output space.

Output orientation

In the case of outputs as categorical variables Equation (5) is the same. The same conclusion as in the case of input orientation can be drawn: A per cannot employ more inputs than the unit under investigation.

The constraint (6) for inputs now reads:

$$\begin{aligned} & x_{ni} - \sum_{j \in J_n'} \lambda_j x_{nj} - \sum_{j \in J_n''} \lambda_j x_{nj} \ \geq 0 \\ & \Rightarrow x_{ni} - \sum_{j \in J_n'} \lambda_j x_{nj} \geq 0 \qquad , n = 1,..,N \end{aligned} \tag{11}$$

where the set J_n' contains peers employing inputs of type n, and the set J_n'' does not. The same conclusions are valid: Peers may have fewer types of inputs than the unit under investigation, but at least must have one type of input in common. One or more types of inputs may be completely missing and thereby reducing the dimensionality of the frontier.

For the outputs in the case of peers having outputs of a type not being produced by the unit under investigation we have the same situation as shown in equation (8), implying that this is possible. But since the equations of this type are inequalities they will not influence the solutions for efficiency score and weights.

In the case of dividing the set into DMUs producing the output of type m and not we have:

$$\begin{aligned}&\sum_{j\in J_m'}\lambda_j y_{mj}-\sum_{j\in J_m''}\lambda_j y_{mj}-\phi_i y_{mi}\geq 0\\&\Rightarrow\sum_{j\in J_m'}\lambda_j y_{mj}-\phi_i y_{mi}\geq 0\quad, m=1,..,M\end{aligned}\tag{12}$$

Again we have, as from (10), that the set J$_m$' cannot be empty. There must be at least one peer producing each of the outputs of DMU$_i$ under investigation.

The inverse of the efficiency score is in the optimal solution calculated from binding constraints:

$$\begin{aligned}&\sum_{j\in J_m'}\lambda_j y_{mj}-\phi_i y_{mi}=0\\&\Rightarrow\phi_i=\frac{\sum_{j\in J_m'}\lambda_j y_{mj}}{y_{mi}}\quad, m\in M'\ ,\end{aligned}\tag{13}$$

where M’ is the set of outputs for which (13) holds with equality. It is only outputs of the type employed by the unit under investigation that will count in the solution for the efficiency measure. If a peer should have more outputs these will not influence the choice of this unit as a peer (i.e. the λ_j - values). The frontier will have the same dimensionality as to outputs as the types produced by the unit under investigation. A peer with fewer outputs than the unit under investigation can compensate, in the expression (13) for an output not produced, for one element less in the sum of the numerator of (13) by higher values for the outputs produced than the other peer DMUs have.

PROPOSITION: Consider a DEA problem with categorical variables in the form of inputs or outputs and at least one input- and one output variable being continuous. Further, the variables are transformed into an exhaustive set of unique types of inputs and outputs, and not all DMUs have a full set of inputs and outputs. Then calculating either an input- or output oriented Farrell efficiency score for a unit implies that:

i) the DMU under investigation will only be compared with peer DMUs having the same or fewer types of inputs

ii) a peer will have at least one type of input in common with the unit under investigation

iii) the DMU under investigation may be compared with peer DMUs having both more and fewer types of outputs, but in the latter case the peer unit must have at least one type of output in common with the DMU under investigation

iv) in the set of peers all types of outputs of the DMU under investigation must be represented.

REMARK: The results for the nature of peers is independent of whether the efficiency measure is input- or output oriented. There is an asymmetry in the results for input and outputs, cf. points i) and iii). This is due to the fact that the variables are constrained to be non-negative, and the inequality constraints for outputs and inputs go in opposite directions, see Models (2) and (3), or equations (5) and (8). More of inputs reduce efficiency while more of outputs improve efficiency. There may be a “bias” against peers having more outputs than the DMU under investigation, because such occurrences do not influence the optimal solution while extra outputs in general draw resources. To overcome this drawback such peers must be “extra” productive. In the same manner, if a peer has fewer outputs than the DMU under investigation, then it has to be especially productive in providing the more limited range of outputs.

4. A note on the literature

Controllable categorical variables in the form of outputs are in Banker and Morey (1986) and Kamakura (1988) only treated as hierarchically ordered, e.g. outputs are classified in categories as “poor” “average”, “good”, or similar orderings, with respect to attributes like quality. The peer DMUs are restricted to be of same or higher service orientation producers. Banker and Morey (1986) obtained the selection of peers from the same or higher ranked groups by changing the objective function in (3) to maximising the sum of (0-1) descriptor binary variables and adding new constraints involving peer group identification, turning the computational problem into one of a mixed-integer LP model. Kamakura (1988) pointed out some inconsistency problems and presented an improved version of the same type of model. The peers were constrained to come only from the same group of DMUs of a higher quality producing DMUs. Kamakura imposed this constraint to avoid forming a reference point by weighing together outputs from different quality classes, as permitted by Banker and Morey.

Rosseau and Semple (1993) reformulated Kamakura's procedure by removing the constraint of peers belonging to the same group, in order to go back to the original objective in Banker and Morey that "..DMUs included in the composite referent point are necessarily of the same or higher service order" (p. 1624). But then mixing peers belonging to different quality groups is accepted. The argument is that obtaining certain levels of outputs with lower ranked inputs, or achieving higher ranked outputs for certain levels of inputs, show a more efficient operation. But the point made in Kamakura (1988) that mix of different qualities has no counterpart in real observations still hangs in the air.

Charnes et al. (1994) acknowledged the criticism of Banker and Morey (1986) in Nakamura (1988), and solved the problem of mixing peers from different quality groups by simply doing away with different quality groups in a special way: For each DMU being investigated the data set is split into two groups: lower quality and same or higher. There is at least one continuous variable associated with each input or output type. The values for the inputs or outputs are simply added together within each of the two groups. The standard DEA LP model is then only run for DMUs belonging to sets in preceding or their own category regarding inputs, to achieve a comparison with DMUs in the same or more disadvantaged categories, or for DMUs belonging to same or higher output quality groups concerning outputs. The same approach is adapted in Cooper et al. (2000) in the case of categorical outputs. The Charnes et al. procedure will then give the same structure of results as in Rosseau and Semple, but with the standard LP formulation and no additional constraints.

If we assume that a single continuous output can be produced in e.g. three different service categories; low, medium and high, we will in our reformulation have three types of outputs measured by the continuous variable. Running an output-oriented DEA efficiency model of type (3) will then imply that DMUs with one service category, low (high) (medium), may be compared with efficient DMUs having this single output category, but also with DMUs having combinations of this category and one or both of the two others. DMUs with a mix of service categories can have peers with different combinations of outputs, e.g. an inefficient unit with two types of outputs, low and medium can have peers with only low outputs, only medium outputs, both low and medium, and both medium and high, low and high or all three types of outputs. Thus, the standard DEA formulation (3) does not conform with the requirement of Banker and Morey that "..DMUs included in the composite referent point are necessarily of the same or higher service order".

To further illustrate the differences between our general formulation and the hierarchical approach in Charnes et al. (1994), the example used in Kamakura (1988) converted to our approach is set out in Table 1.

Table 1. Three output quality categories

Unit	Input	Output high	Output medium	Output low
A	1	2	0	0
B	5	4	0	0
C	2	0	0	2
D	4	0	3	0
E	1	0	1	0

The consequences of using the Charnes et al. (1994) procedure by ordering the output categories from least to most desirable and selecting DMUs from the sets with the same or higher ordering, can be illustrated using the data in Table 1. The unit with low quality may be compared with DMUs from all the three quality groups, there are no restrictions on the choice of peers or how reference points are related to the peers. The three output columns in Table 1 are simply added together. For the case of medium quality DMUs the peers can either have medium or high quality implying that the reference point on the frontier can be a mix of medium and high quality DMUs, the output columns for "Medium" and "High" in Table 1 are added together. For high quality DMUs this group has to be run separately.

Kamakura (1988) used a variable returns to scale specification of the DEA model (as did Banker and Morey, 1986 and Charnes et al., 1994). This is achieved by adding the following constraint to the DEA model (3):

$$\sum_{j=1}^{J} \lambda_j = 1 \tag{14}$$

The results from running model (3), with (14) as additional constraint in the general mode with three outputs, and doing the hierarchical ordering as in Charnes et al., are set out in Table 2. Our general model yields *all* DMUs efficient and each unit only compared with DMUs having the same type of outputs. The apparently inefficient DMUs B and D in the high quality and medium quality groups respectively becomes technically efficient (but not scale efficient) due to the models really having only two dimensions, and no unit dominating the other. Figure 1 illustrates the three frontiers for the groups, the solid line between DMUs A and B, and the vertical and horizontal extensions from A and B, constitute the frontier for the high quality group, the broken line between DMUs E and D, including the vertical and horizontal extensions from the points E and D, constitute the frontier for the medium quality group, and the vertical and horizontal broken lines out from point C form the frontier for the low quality group of one unit.

In general, in the case of DMUs having outputs only of one service category, as is the case in Banker and Morey (1986) and Kamakura (1988), each unit will be compared with DMUs of only the same category according to point ii) or iii) of the proposition above. But this is the same as running *separate* DEA problems for each group.

Table 2. Results for the standard VRS model [(3) and (14)], and Charnes et al.(1994) ordering

Unit	Efficiency score model (3)	Efficiency score for {high}	Efficiency score for {medium+high}	Efficiency score for {low+medium+high}
A	1	1	1	1
B	1	1	1	1
C	1	-	-	0.8
D	1	-	0.86	0.86
E	1	-	0.5	0.5

The Charnes et al. procedure treats the highest quality DMUs as one group and we have the same results as in the standard DEA LP model for the high quality DMUs. Considering high and medium quality DMUs we get DMUs D and E as inefficient and the high quality DMUs being peers. Aggregating the continuous output for all quality groups in Table 1 we get the efficiency scores in the last column in Table 2, and the low quality unit appears as inefficient even disregarding the information that its output is of lowest quality. Since the peers remain the same the efficiency scores for the medium quality DMUs are replicated. The frontier for all DMUs together is formed by the line between A and B and the vertical – horizontal extensions (the former going through E).

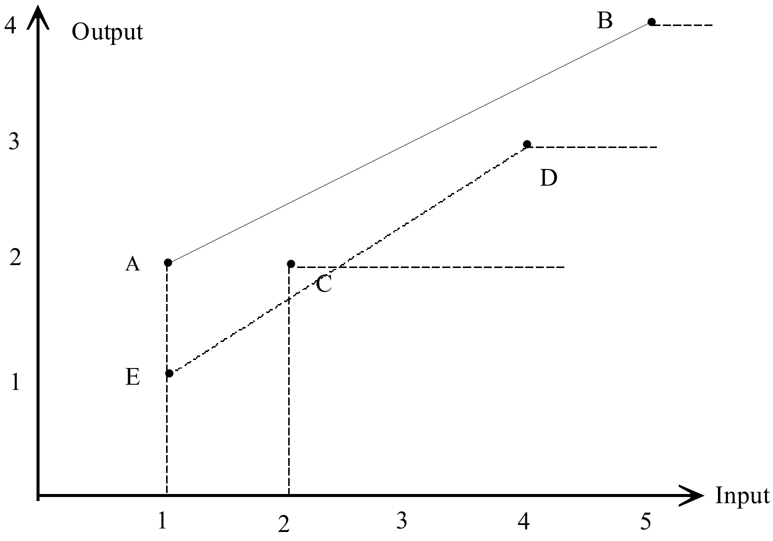

Figure 1. The frontiers

We may also consider inputs as in Charnes et al., and take as example input (labour) being of low, medium and high quality, and hours used as the continuous variable. Then we will in the standard model (2) have three input types all measured by hours as unit. Assuming one or more continuous outputs, and running either an input- or output oriented efficiency model, a unit using only labour of type low (high) (medium) quality can only be compared with DMUs using low (high) (medium) quality, whereas a unit with labour in all categories can be compared with DMUs using from all the three types to only one type (be it of low, medium or high quality). In the case of all DMUs employing only a single type of input, running separate DEA models for each group gives identical results. (One does not have to separate the groups before running a standard DEA programme, the solution will automatically have this separation property.)

In Puig - Junoy (1998) it is stated that with respect to the hierarchical categorical input, probability of survival at the time of hospitalisation, one is only interested in comparing DMUs that employ the same types of inputs (p. 268). The model used is Charnes et al. (1994). However, our standard model can easily accommodate the wish because the categorical input can only be in one of three states for each DMU. Separate DEA models may also, as well, have been run.

5. Conclusions

Categorical variables are often used when performing efficiency analyses of multiple input- multiple output production units. A hierarchical structure has so far been imposed in the literature, leaving us with a mixed integer LP programme if mix of differently ranked peers is not allowed. A standard LP format of the DEA model can be used if a special aggregation of types into two groups is done; according to lower or higher ordered types compared with the situation for the DMU under investigation. The procedure in Charnes et al. (1994) implies that all differences between types of variables can be reduced to two, and the DEA model is only run for the one subset relevant for the analysis. But then there are no restrictions on the mixing of peers with lower ranked inputs, or higher ranked outputs respectively. It is implicitly assumed that it makes sense to mix qualities in both these subsets. Without a hierarchical ordering the formation of two subsets is not defendable, and with an ordering one may, as in Kamakura (1988), question the procedure because mixes cannot be observed.

Our approach is designed for situations when it is not natural to order categorical variables hierarchically. A standard LP format of the DEA model can be used, if both categorical and continuous variables are present, by writing out all the combinations of the categorical variables as different types of inputs and/or outputs. Most DMUs will then not have full sets of positive variables. Using a standard LP DEA model of type (2) or (3) will then not in general give the same results (with respect to efficiency scores and peers) as using the mixed integer-LP model of Banker and Morey (1986) or the reformulation in Kamakura (1988), or the special aggregation introduced in Charnes et al. (1994).

We have formulated a more general setting with no ordering of categories and investigated the nature of the selected peers both in the input- and output dimension. The special cases dealt with in the literature can easily be incorporated. If each DMU has only one of the possible types of inputs or outputs, and a comparison only with same or higher ranked types is wanted, formation of new subsets of DMUs as required in Charnes et al. (1994), is not necessary, because employing the standard model with the full set of types of variables will yield separate group results for DMUs with same type of variable by definition.

The general result for the characterisation of peers is that there is a basic asymmetry between inputs and outputs due to the inequality constraints going in opposite directions and all variables restricted to being non-negative. More of inputs reduce efficiency while more of outputs improve efficiency. A peer may have at most the same types of inputs, but may have less than the DMU under investigation, but may have either fewer or more outputs. There may thus be a mix of peers with different characteristics than the DMU under investigation, but a peer must always have at least one input, and at least one output, in common. Notice that such mix of peers may also be observable in principle.

References

Banker, R. D. and R. C. Morey (1986): "The use of categorical variables in Data Envelopment Analysis", *Management Science* 32(12), 1613-1627.

Charnes, A., W. W. Cooper, A. Y. Lewin and L. M. Seiford (eds.): *Data Envelopment Analysis: Theory, Methodology, and Applications*, Boston/Dordrecht/London: Kluwer Academic Publishers, 1994, Section 3.3 Categorical inputs and outputs, pp. 52-54.

Cooper, W. W., L. M. Seiford, and K. Tone (2000): *Data Envelopment Analysis. A comprehensive text with models, applications, references and DEA-solver software*, Boston/Dordrecht/London: Kluwer Academic Publishers, Section 7.4 DEA with categorical DMUs, 193-197.

Erlandsen, E. and F. R. Førsund (2001): “Efficiency in the Provision of Municipal Nursing- and Home-Care Services: the Norwegian Experience”, in K. J. Fox (ed.): *Efficiency in the public sector*, Boston/Dordrecht/London: Kluwer Academic Publishers, forthcoming.

Førsund, F. R. and R. Zanola (2001): "Selling Picasso paintings: The efficiency of auction houses", *ICER Working Paper* No 7/2001.

Kamakura, W. A. (1988): "A note on the use of categorical variables in Data Envelopment Analysis", *Management Science* 34(10), 1273-1276.

Puig - Junoy, J. (1998): "Technical efficiency in the clinical management of critically ill patients", *Health Economics* 7, 263-277.

Rousseau, J. J. and J. Semple (1993): “Categorical outputs in Data Envelopment Analysis,” *Management Science* 39(3), 384-386.