International Journal of Business and Economics, 2002, Vol. 1, No. 1, 33-43

Categorical Variables in DEA

Finn R. Førsund[*]
Department of Economics, University of Oslo, Norway

Abstract

The standard DEA model can be applied to a mix of categorical and continuous variables by entering all combinations of them as different types of inputs and/or outputs. Theoretical implications for the nature of feasible peers are investigated.

Key words: categorical variable; DEA; efficiency; linear programming; peer

JEL classification: C61; D24

1. Introduction

In efficiency analysis of production units (DMUs) the input and output variables are usually assumed to be continuous. 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 Data Envelopment Analysis (DEA) applications, especially for DMUs where outputs are not sold on markets, that output variables are categorical, e.g., court cases completed are categorised into civil cases, assault, robbery, economic crimes, etc. Quality aspects will often be of the categorical kind. But inputs may also be categorical, as is labour with different types of education.

It should be noted that a general assumption underlying the rationale for comparing different production units by calculating efficiency scores is that the inputs and outputs are indeed comparable, i.e., that they are homogeneous across DMUs. Labour input measured in 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. One way to ensure comparability is to form categorical variables.

DEA models with categorical variables are treated for the first time (to my

Received June 25, 2001, accepted July 27, 2001.

[*]Correspondence to: Department of Economics, University of Oslo, Box 1095, 0317 Blindern, Oslo, Norway. Email: f.r.forsund@econ.uio.no. The paper was written while being a 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. Constructive comments from a referee are gratefully acknowledged.

knowledge) in Banker and Morey (1986), and the approach is 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 Puig-Junoy (1998) and also presented in Cooper, Seiford, and Tone (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 types "low," "medium," and "high" quality). 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 variables only. A crucial assumption needed is that there is at least one continuous input 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 (2002) use a limited set of age groups as categorical output variables and the number of clients within each group as the continuous variable, as well as whether nursing homes have single rooms or not. In a study of the efficiency of auction houses in 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 most often do not have observations of all variables, it is of interest to study the nature of peers with respect to composition of variables. The situation with hierarchically ordered categorical variables comes 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. An illustration of the different approaches is provided in Section 4, and some concluding remarks are 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:

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

where x is the input vector and y is the output vector, and in the last expression we have introduced J observations and indexed output by m and input by n. The variables $\lambda_j (j = 1, ..., J)$ are non-negative weights (intensity variables) defining frontier points. Constant returns to scale is assumed for simplicity. The nature of scale does not matter for the question of model specification type with categorical

variables. Basic standard properties are that the production set is convex and includes all points and that envelopment is done with minimum extrapolation, i.e., the fit is as "tight" as possible.

The input- and output-oriented Farrell radial efficiency measures, respectively E_{1i} and E_{2i} for each DMU i (henceforth DMU_i) in the set of J observations, are calculated by solving the following linear programmes set up according to the definition of the measures (with the necessary change that in the output-oriented case we solve for the inverse measure $\phi_i = 1/E_{2i}$ in order to maintain a linear programming problem):

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

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

Each type of input is scaled down with the same factor, θ_i, and each type of output is scaled up with the same factor, ϕ_i, until the frontier is reached according to the definition of the Farrell efficiency measures. DMUs with positive λ_i (for convenience the same symbol is used in the input- and output- oriented cases) 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 input (output) constraints, the radial contraction (expansion) of the DMU_i observation coincides with the weighted peer values as the comparison point.

3. Features of the DEA Solution with Categorical Variables

3.1 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 $\text{DMU}_j\,(j=1,...,J)$ 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 a continuous input variables of type $n\,(n=1,...,N)$ and y_{mj} be a continuous output variables of type $m\,(m=1,...,M)$. We then have $K \times N$ different types of inputs (each continuous input variable is matched with each of the K types of inputs) and $P \times M$ different types of outputs (each continuous output variable is matched with each of the P types of categorical outputs). Thus, the situation with a mix of categorical and continuous variables is converted to a standard DEA LP model. Note that we may run into a dimensionality problem as to the number of observations and number of variables. One way out is to use some categorical variables as variables in a second stage of correlating efficiency scores with explanatory variables (e.g., type of ownership).

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

3.2 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 by at least one unit. The production unit under investigation is DMU_i. The restriction system for inputs is:

$$\theta_i x_{ni} - \sum_{j=1}^{J} \lambda_j x_{nj} \geq 0,\ n=1,...,N\,. \tag{4}$$

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. Fulfillment of the constraint then requires all the products of λ_j and x_{nj} to be zero. If x_{nj} is positive for a peer j (i.e., we are considering an input type that 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 that the unit under investigation is not using [see also 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 as J'_n and the set of DMUs without it as J''_n. The constraint then reads:

$$\theta_i x_{ni} - \sum_{j \in J'_n} \lambda_j x_{nj} - \sum_{j \in J''_n} \lambda_j x_{nj} \geq 0 \;\Rightarrow\; \theta_i x_{ni} - \sum_{j \in J'_n} \lambda_j x_{nj} \geq 0, n = 1, ..., N \;. \quad (6)$$

The x_{nj} $(n = 1, ..., N)$ 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 may be calculated from a binding constraint of type (6):

$$\theta_i x_{ni} - \sum_{j \in J'_n} \lambda_j x_{nj} = 0 \;\Rightarrow\; \theta_i = \frac{\sum_{j \in J'_n} \lambda_j x_{nj}}{x_{ni}}, \; n \in N', \quad (7)$$

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:

$$\sum_{j=1}^{J} \lambda_j y_{mj} \geq 0 \;. \quad (8)$$

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, then 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 as J'_m and the set of DMUs without it as J''_m. The constraint then reads:

$$\sum_{j\in J'_m} \lambda_j y_{mj} - \sum_{j\in J''_m} \lambda_j y_{mj} - y_{mi} \geq 0 \;\Rightarrow \sum_{j\in J'_m} \lambda_j y_{mj} - y_{mi} \geq 0, m = 1, ..., M \,. \tag{9}$$

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 $M \times N$ 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.

3.3 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 peer cannot employ more inputs than the unit under investigation. The constraint (6) for inputs now reads:

$$x_{ni} - \sum_{j\in J'_n} \lambda_j x_{nj} - \sum_{j\in J''_n} \lambda_j x_{nj} \geq 0 \Rightarrow x_{ni} - \sum_{j\in J'_n} \lambda_j x_{nj} \geq 0, n = 1, ..., N \quad , \tag{11}$$

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 must have at least one type of input in common. One or more types of inputs may be completely missing and thereby reduce 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 those which do not, we have:

$$\sum_{j\in J'_m}\lambda_j y_{mj} - \sum_{j\in J''_m}\lambda_j y_{mj} - \phi_i y_{mi} \geq 0 \Rightarrow \sum_{j\in J'_m}\lambda_j y_{mj} - \phi_i y_{mi} \geq 0,\ m = 1,...,M. \quad (12)$$

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:

$$\sum_{j\in J'_m}\lambda_j y_{mj} - \phi_i y_{mi} = 0 \Rightarrow \phi_i = \frac{\sum_{j\in J'_m}\lambda_j y_{mj}}{y_{mi}},\ m \in M', \quad (13)$$

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 or fewer types of outputs, but 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.

***Remarks*:** 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 (2) and (3) or equations (5) and (8). More inputs reduces efficiency while more 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. This asymmetry may be of help for the classification of variables into inputs and outputs.

4. Illustration of the Approaches

Controllable categorical variables in the DEA literature are only treated as hierarchically ordered, e.g., classified into categories as "poor," "average," and "good," or similar orderings, with respect to attributes like quality. The peer DMUs are restricted to be of same or higher service orientation producers. There is a discussion whether peers should belong to only one type, as advocated by Kamakura (1988), or a mix of types, as in Banker and Morey (1986) and reformulated by Rosseau and Semple (1993). Charnes et al. (1994) solved the problem of mixing peers from different quality groups by simply doing away with different quality groups in a special way. Remember that there is at least one continuous variable associated with each input or output type. For each DMU being investigated the data set is split into two groups: DMUs belonging to higher quality groups, or lower and their own category regarding inputs, and DMUs belonging to the same or higher output quality groups, or lower groups concerning outputs. To achieve a comparison with DMUs in the same or more disadvantaged categories, the standard DEA LP model is then only run for DMUs belonging to the set of DMUs in the lower or their own quality category regarding inputs, and for DMUs belonging to the same or higher output quality groups concerning outputs. All feasible sets are covered.

The same approach is adapted in Cooper, Seiford, and Tone (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.

To illustrate the differences between our general formulation and the hierarchical approaches in Kamakura (1988) and in Charnes et al. (1994), the five data points used in Kamakura used for constructing frontiers are set out in Figure 1 with the quality indications H, M, and L. As a variable returns to scale is used in the literature, we will add to problem (3) the constraint that the weights, λ_j, add to one.

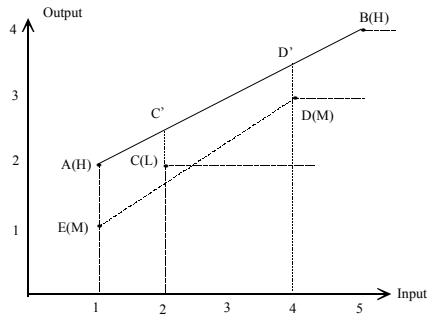
According to the Charnes et al. procedure, the unit with low quality may be compared with DMUs from all the three quality groups; there are no restrictions on

the choice of peers or how reference points are related to the peers. The frontier is the connection line between A and B with the vertical extension through E and the horizontal extension from B. Both the medium-quality DMUs and the low-quality DMU become inefficient. 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. For high-quality DMUs, this group has to be run separately. The frontiers for all three runs remain the same. Thus the peers for all three possible groups remain units A and B.

The Kamakura procedure happens to yield the same frontier and the same efficiency scores for the data introduced by Kamakura. But now it is only the high-quality DMUs A and B that can serve as peers in principle, because of the requirements that they have to be from the same or higher quality group and that they can only belong to one group.

Our general model yields three different frontiers and all DMUs efficient, i.e., all units are peers within their quality groups. Since the DMUs have only one type of output, each DMU is only compared with DMUs having the same type of outputs. The data set disaggregates to three sets without any interaction between them. The inefficient DMUs E and D in the medium-quality group and C in the low-quality group become technically efficient (but not scale efficient). 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. The difference between the hierarchical approaches and our approach is that the former allows peers to be from different quality groups, while the latter demands at least one common type.

Fig. 1. The frontiers



In general, in the case of DMUs having outputs only of one service category, as 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 as illustrated above. Similarly, 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, our 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). Our standard model, however, accommodates this because the categorical input can only be in one of three states for each DMU. Separate DEA models may also have been run.

5. Conclusions

When dealing with categorical variables in DEA models, a hierarchical structure has so far been imposed in the literature. 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 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), 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 in both the input and output dimensions. A general feature of 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. 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 peers with a different mix of characteristics than the DMU under investigation, but a peer must always have at least one input, and at least one output, in common. The results in Førsund and Zanola (2001) illustrate the empirical importance of mix of characteristics and linkage effects through types in common with peers and the DMU under investigation.

The special case of Charnes et al. (1994) can easily be incorporated. If each DMU has only one of the possible types of inputs or outputs, and a comparison only with the 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 the same type of variable by definition.

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, (1994), *Data Envelop-*

ment Analysis: Theory, Methodology, and Applications, Boston/Dordrecht/ London: Kluwer Academic Publishers, [Section 3.3 Categorical Inputs and Outputs], 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, (2002), "Efficiency in the Provision of Municipal Nursing- and Home-Care Services: the Norwegian Experience," in *Efficiency in the public sector*, K.J. Fox, ed., Boston/Dordrecht/London: Kluwer Academic Publishers, 273-300.

Førsund, F. R. and R. Zanola, (2001), "Selling Picasso Paintings: The Efficiency of Auction Houses," *ICER Working Paper* No. 7. (www.icer.it)

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.