MODELING BLANK DATA ENTRIES IN DATA ENVELOPMENT ANALYSIS

TIMO KUOSMANEN[1]
Wageningen University
Department of Social Sciences
Hollandseweg 1
6706 KN Wageningen
The Netherlands
E-mail: Timo.Kuosmanen@wur.nl
Tel: +31 317 484 738
Fax: +31 317 484 933

ABSTRACT

We show how Data Envelopment Analysis (DEA) can handle missing data. When blank data entries are coded by appropriate dummy values, the DEA model automatically excludes the missing data from the analysis. We extend this result to weight-restricted DEA models by presenting a simple modification to the usual weight restrictions, which automatically relaxes the weight restriction in case of missing data. Our approach is illustrated by a case study, describing an application to international sustainable development indices.

Key Words: *Data Envelopment Analysis, Weight Restrictions, Missing Data, Blank Entries*

1. Introduction

Missing data are a chronic disease in applications of Data Envelopment Analysis (DEA: Charnes, Coper, and Rhodes, 1978). Very often, potentially important input and/or output variables have insufficient coverage, or Decision-Making Units (DMUs) fail to report all required statistics. The art of composing the dense data matrices required by DEA from the sparse statistics available is a critical step of the analysis in a wide variety of applications.

[1] This research forms part of the research program on Nonparametric Methods in Economics of Production, Natural Resources and the Environment. See http://www.sls.wageningen-ur.nl/enr/staff/kuosmanen/program1/ for further information. I am grateful to Larry Cherchye for inspiring discussions. The financial support from the Emil Aaltonen Foundation, Finland, is gratefully acknowledged.

Due to its nonparametric and multi-dimensional nature, DEA approach generally requires large numbers of DMUs to produce statistically meaningful results. (See e.g. Simar and Wilson, 2000, for discussion of statistical properties of DEA efficiency estimators.) Therefore, DEA is highly vulnerable to the data problems. Still, the treatment of missing data has only attracted some passing remarks in the literature. One exception is Kao and Liu (2000), who use fuzzy sets to model the ranges for missing data. Yet, eliminating blank entries from the data matrices remains the most standard approach of handling the problem of missing data.

In this paper we wish to challenge the conventional wisdom that eliminating the blank entries from the input-output data matrices would always be necessary. By eliminating the blank entries we inevitably discard entire rows or columns of the data matrices, and hence are bound to lose lots of valuable information of the production possibilities. If the blank entries do not cause any other harm to the analysis, besides the missing information of the specific data entry itself, then discarding existing, available information certainly cannot improve the matters. The main challenge of this paper is therefore to show how incomplete data, containing blank entries, can enter the DEA model without influencing the efficiency measures. We show that in the basic DEA models, representing the blank entries by appropriately chosen dummy numbers is equivalent to *excluding* the missing input or output from the calculation of the efficiency score of the corresponding DMU. This result does not depend on the orientation of measurement or the returns-to-scale specification, and hence applies to most DEA models.

For outputs, using zero as a dummy for blank entries proves to be an effective solution. The question of blank output entries is hence closely related to the treatment of zeros in the data matrices (see e.g. Thompson et al, 1993, for discussion). The classic article of Charnes, Cooper and Rhodes (1978) required all input-output data of DMUs to be *strictly positive*, and hence left no room for blank entries. The subsequent research and discussions have further elaborated on the *minimal* data requirements, and considerably relaxed the positivity condition. The minimal conditions are now known to be: 1) At least one DMU consumes/produces every input and output,

2) Each DMU consumes at least one input and produces at least one output (Färe and Grosskopf, 2002; following Shephard, 1970). There are therefore no technical obstacles for using zeros as dummies for missing data in the output matrices.

Another closely related topic concerns modeling additional weight restrictions in DEA. The weight restrictions have established in the DEA applications as a standard tool for increasing the discriminatory power of the model by incorporating additional information of the technical or value tradeoffs between inputs and/or outputs (see e.g. Allen et al., 1997, for a review). When the data includes blank entries, however, one should be careful that the arbitrary dummies for missing data do not influence the results because of the weight restrictions. To remedy this problem, we present a simple modification of the standard weight restriction, which functions normally for the observed data, but relaxes the weight restriction in case of dummies for blank entries so that the missing data will not count in the analysis.

The remainder of the paper is organized as follows. In Section 2 we demonstrate how the basic DEA models can accommodate missing entries, and automatically exclude them from the analysis. Section 3 presents a simple procedure for relaxing weight restrictions in cases of missing entries. To illustrate the practical usefulness of these results, Section 4 reviews (as a case study) the empirical application to international sustainable development benchmark indices from which the model developments of this paper originally stem from. Section 5 draws the concluding remarks, and illustrates the usefulness of these model developments by discussing the empirical application that gave us the first inspiration to investigate these issues.

2. Modeling missing entries in DEA

The dimensions of our DEA model are n DMUs, r inputs, and s outputs. Adopting the standard notation, X denotes a $n \times r$ input matrix, and Y denotes $n \times s$ output matrix, respectively. We assume both matrices are non-negative, and every row and column includes at least one strictly positive element.

Adhering to the multiplier side formulation, we can solve the standard radial DEA efficiency scores of the evaluated DMU *k* as:

Input-oriented model

$$\max_{u,v,w} \phi_k = \sum_{j=1}^{s} Y_{kj} u_j + w$$

s.t.

$$\sum_{j=1}^{r} X_{kj} v_j = 1;$$

$$\sum_{j=1}^{s} Y_{ij} u_j - \sum_{j=1}^{r} X_{ij} v_j + w \le 0 \ \forall i = 1,...,n$$

$$u_j, v_j \ge 0$$

Output-oriented model

$$\min_{u,v,w} \theta_k = \sum_{j=1}^{r} X_{kj} v_j + w$$

s.t.

$$\sum_{j=1}^{s} Y_{kj} u_j = 1;$$

$$-\sum_{j=1}^{s} Y_{ij} u_j + \sum_{j=1}^{r} X_{ij} v_j + w \ge 0 \ \forall i = 1,...,n$$

$$u_j, v_j \ge 0$$

***Returns to Scale (both models)*:**

$VRS : w$ free
$NDRS : w \ge 0$
$NIRS : w \le 0$
$CRS : w = 0$

For simplicity, we leave the non-radial slacks outside this discussion.

Suppose for one reason or another, data of output *j* for the evaluated DMU *k* is missing, i.e., Y_{kj} is a blank entry. There are two basic alternatives for eliminating DMU *k* from the data set, or perhaps still worse, discarding a potentially highly relevant output *j* from the entire DEA analysis:

1) Omit output *j* from the calculation of the efficiency score for DMU *k* (but keep it available for other DMUs in the sample).
2) Insert a dummy value $Y_{kj} = 0$ in the output matrix.

Theorem 1: Approaches 1) and 2) above are equivalent, yielding equal DEA efficiency scores.

Proof. Suppose we set $Y_{kj} = 0$. Since by assumption DMU *k* has produced a strictly positive amount of at least one output, say output *i*, the sum $\sum_{j=1}^{s} Y_{kj} u_j$ always increases as we shift some weight from output *j* to output *i*. Thus, the optimal output weights will always satisfy $u_j^* = 0$. Consequently,

output *j* does not count in the efficiency index, that is, both approaches exclude output *j* from the calculation of the efficiency index. □

From the applied perspective, the latter procedure is much more convenient, since it does not involve any DMU-specific modifications to the computation code. Even though the zero value merely represents a dummy for the blank entry, we have a meaningful interpretation of the model: We have the same result as if we had solved the efficiency score using the output measures present in the data set.

Remarkably, the result does not depend on the (input/output) orientation of measurement. A similar trick also applies for the missing inputs. Instead of using the dummy value of zero, however, we should use some sufficiently large number, say $M >> \max_{h,j}\{X_{hj}\}$. A simple way of testing whether *M* is large enough is to check from the optimal solution whether $v_j^* = 0$ for all $X_{kj} = M$.

3. Missing Entries and Weight-Restrictions

The treatment of the previous section basically exploited the possibility to assign zero weight to output and input dimensions in which the DMU is performing poorly, which is equivalent to excluding those dimensions from the analysis. In many applications, however, we are inclined to think that DMUs should not be able to hide away their poor results in some performance dimensions by simply assigning all weight to other dimensions, but rather, all data should count. It is nowadays a standard and widely used approach to impose additional external restrictions on the weight flexibility (see e.g. Allen et al., 1997, for a review of techniques). Since the input weights are scaled according to the evaluated DMU such that $\sum_{j=1}^{r} X_{kj} v_j = 1$ in the input-oriented models, and the output weights are scaled such that $\sum_{j=1}^{s} Y_{kj} u_j = 1$ in the output-oriented models, restricting the

absolute levels of weights *u* and *v* does not typically make much sense. Therefore, the most standard way of modeling these restrictions is to impose constraints of ratio form

$$a_{hi} \le \frac{u_h}{u_i} \le b_{hi}, \quad 0 < a_{hi} < b_{hi} < 1 \tag{1}$$

which characterize the feasible range for the weight of output *h*, compared to the weight of output *i*. A weight-restricted DEA model is obtained by imposing constraints of type (1) [in the linearized form] in the basic DEA models presented in the previous section.

A problem we face when using dummy values 0 or *M* for missing outputs and inputs, respectively, in connection with the weight-restrictions, is that weight restrictions can nullify the equivalence result of Theorem 1. Not only the additional weight restrictions limit DMU's freedom to completely discard some unwanted input-output variables, also the arbitrary dummies for blank entries will be assigned a strictly positive weight.

To discard the blank entries from the analysis, we would like to impose restrictions of type (1) on those outputs which are actually observed, but relax this weight restriction in case of missing entries to preserve the equivalence result of Theorem 1. Formally, this could be modeled by writing the weight restriction as a *disjunctive constraint*

$$a_{hi} \le \frac{u_h}{u_i} \le b_{hi}, \quad \text{if } h,i \notin D, \tag{2}$$

where *D* is the index set of the missing entries in the output matrix *Y*. Obviously, we could resort to simple heuristics and simply modify the constraints for each DMU and each output separately. This approach would work fine, but it is a frustratingly time consuming procedure in large applications when there are many missing entries pertaining to different outputs.

The following theorem presents a simple but effective trick for modeling the disjunctive constraint (2) using simple linear inequalities.

Theorem 2: The disjunctive weight restriction (2) can be equivalently modeled as a pair of linear inequalities

$$\begin{aligned}(u_h - a_{hi}u_i)\cdot(Y_{ki}\cdot Y_{kh}) \geq 0\\ (u_h - b_{hi}u_i)\cdot(Y_{ki}\cdot Y_{kh}) \leq 0\end{aligned} \quad (3)$$

Proof. Reorganizing constraint (1), we have the inequalities $(u_h - a_{hi}u_i) \geq 0, (u_h - b_{hi}u_i) \leq 0$. To model the disjunctive part of (2), we simply multiply both sides of the inequalities by the product $Y_{ki}\cdot Y_{kh}$, which is an exogenously given constant. Clearly, if output *h* or *i* is missing, then $Y_{ki}\cdot Y_{kh} = 0$, and the inequalities (3) hold for all weights *u*. However, when $Y_{ki}\cdot Y_{kh} > 0$, inequalities (3) characterize the same weight range as inequalities (1). □

This result does not depend on the orientation of measurement or the returns to scale specification. It also extends to the missing inputs. Recall from the previous section that we label the blank entries of the input matrix by some sufficiently large number $M >> \max_{h,j}\{X_{hj}\}$. We can model a disjunctive constraint of type (2) for the input weights *v* by we write the following pair of linear inequalities:

$$\begin{aligned}(v_h - a_{hi}v_i)\cdot(M - X_{ki})\cdot(M - X_{kh}) \geq 0\\ (v_h - b_{hi}v_i)\cdot(M - X_{ki})\cdot(M - X_{kh}) \leq 0\end{aligned} \quad (4)$$

In (4), the left-hand sides of the inequalities will be zero, and the inequalities become redundant, whenever input *h* or *i* is assigned the dummy value *M*. However, the usual weight restriction is invoked whenever the input levels coincide to the observed range. The usefulness of these modeling developments will become evident when considering the following empirical case.

4. Case Study: International Benchmarks in Sustainable Development

The DEA literature has traditionally been highly application oriented in the sense that motivation for the new model developments typically stems from problems or limitations experienced in empirical application of the method. The present study is not an exception. The need for these model developments arised in the Sustainable Development (SD) study, reported in more detail in

Cherchye and Kuosmanen (2002). In that study we explored the possibilities of using DEA to construct a so-called Meta-index of Sustainable Development (MISD), which would serve as a tool for international benchmarking. To identify the most developed benchmarks at different income levels, we ranked countries according to the weighted-average of 14 different aggregate indices of SD and its economic, social/political, and environmental sub-components, reported by well-established international organizations (like the United Nations Development Program) or distinguished scientific expert teams. Treating the normalized values of different indices as outputs, we derived the unequal weights of the meta-index using the DEA approach.

The practical problem we faced in applying DEA in this setting was, the full data of 14 output measures we wanted to use was only available for 15 countries. Therefore, excluding all countries with missing data was not an option. We decided to include all countries with the minimum of 6 outputs in our data set to increase the sample size up to 154, by evaluating each country only in terms of those performance dimensions for which data existed. This seemed a reasonable approach in our *meta*-level assessment, given that each underlying SD index (=output in the DEA model) should already constitute a well-balanced aggregate view of SD as such.

As a result, our data set took the form of a 14×154 output matrix. This sparse output matrix included 395 blank entries, which amount to 18 per cent of the total of 2156 elements of the matrix. On the average, there were 2.56 missing entries for each country (mode = median = 2). Considering the size of the problem, it would have taken tremendous effort to make DMU-specific modifications to the computation code to ensure that each DMU is evaluated using only those outputs for which the data was available. Using Theorem 1, we could simply replace the blank entries by zeros, and execute the standard DEA code. (The GAMS code we used in this SD study is available in the Appendix.)

Given the large number of output dimensions, it was also necessary to reduce weight flexibility. By trial and error, we realized that the weight restrictions interfere with the attractive interpretation of Theorem 1, in other words, our zero-valued blank entries are assigned a positive

weight in our index. To eliminate the influence of the arbitrary dummies for blank entries, we first considered revising the problem formulation and the computer code for each 154 countries separately, to eliminate the weight restrictions from the missing zero entries. However, we soon discovered this would take enormous amounts of time and effort. This motivated us to look for a simple but effective techniques to lift weight restrictions from the zero entries in automated fashion. That investigation eventually paid off in the form of the result proved in Theorem 2.

It is difficult to estimate the effective time saving facilitated by the model developments presented in the previous sections. It seems likely that we would never have completed the study, had we not found these simple but effective tricks for dealing with blank entries. By sharing these results, we hope to improve the application possibilities of DEA to similar application situations plagued by blank entries.

5. Concluding remarks

In many applications, the input and output matrices include blank entries, which are eliminated before the DEA analysis. Still, DMUs with missing data could be highly useful as reference or benchmarks units, which span the efficient frontier. Since DEA builds up virtual DMUs as convex combinations of observed reference DMUs, including the missing entries will not deteriorate the empirical production possibilities frontier. By contrast, those DMUs for which it is difficult to obtain the full input-output data might represent desirable ‘diversity’, reflected in the wider spread of input-output mixes for instance.

In this paper we noted that labeling the missing entries with appropriate dummy variables (zero for missing outputs; some sufficiently large number for inputs), we can run the normal DEA model automatically in such a way that the blank entries do not count. Another new trick presented in this paper was a simple modification of the weight restrictions to preserve this property. More precisely, we considered a disjunctive weight restriction that automatically relaxes the weight restriction in case of missing values, and presented a simple but effective trick to linearize this

constraint. We think these insights can tremendously increase our capabilities of applying DEA in situations where data coverage presents problems.

As a final note, we see that it can be unfair for DMUs with missing entries to be included in the efficiency rankings or productivity comparisons with zero outputs or arbitrarily large inputs when the true performance is better than that. Since we assume the most pessimistic values for the missing data, DMUs which fail to report all input-output variables are generally handicapped in comparisons to DMUs whose outputs are correctly represented. In many cases, however, it seems fair to reward DMUs that openly report their data, by assessing them in terms of a larger number of input-output dimensions. The case study reviewed in Section 4 is a good example. This can encourage DMUs to pay more attention on reporting data in the future, an important consideration in international comparisons in particular. The strategic dimensions of how the treatment of missing data in efficiency assessment gives the correct incentives to report truthful data deserves further research.

References

Allen, R., A. D. Athanassopoulos, R.G. Dyson and E. Thanassoulis (1997): Weight Restrictions and Value Judgements in DEA: Evolution, Development and Future Directions, *Annals of Operations Research* 73, 13-34.

Charnes, A., W.W. Cooper, and E. Rhodes (1978): Measuring the Efficiency of Decision Making Units, *European Journal of Operational Research*, 2(6), 429 – 444.

Cherchye, L., and T. Kuosmanen (2002): Benchmarking on Sustainable Development: A Synthetic Meta-Index Approach, unpublished working paper. Available on-line at: http://www.sls.wau.nl/enr/staff/kuosmanen/papers/MISD.pdf

Färe, R., and S. Grosskopf (2002): Two Perspectives on DEA: Unveiling the link between CCR and Shephard, *Journal of Productivity Analysis* 17(1-2), 41-47.

Kao, C., and S.-T. Liu (2000): Data envelopment analysis with missing data: an application to University libraries in Taiwan, *Journal of the Operational Research Society* 51(8), 897-905.

Shephard, R.W. (1970): *Theory of Cost and Production Functions*, Princeton University Press.

Simar, L. and P. Wilson (2000): Statistical Inference in Nonparametric Frontier Models: The State of the Art, *Journal of Productivity Analysis* 13(1), pp. 49-78.

Thompson R.G., P.S., Dharmapala, and R.M. Thrall (1993): Importance for DEA of Zeros in Data, Multipliers, and Solutions, *Journal of Productivity Analysis* 4, 379-390.

Appendix

GAMS code used in the case study of Section 4.

```
SETS
i     performance criteria /i1*i14/
j     countries /j1*j154/;

alias(j,k);
alias(i,l);

PARAMETERS
Y(i,j)              the value of performance indicator i for country j ;

VARIABLES
TOTOBJ              total objective
OBJ(j)              objective of country j

POSITIVE VARIABLES
W(i,j)              weights (or price) of indicator i for country j;

EQUATIONS
QTOTOBJ             total objective
QOBJ(j)             equation for objective of contry j
QKONST(j,k)         constraint
QWR(i,j,l)          weight restriction

QOBJ(j).. OBJ(j)=l=sum(i,W(i,j)*Y(i,j));
QTOTOBJ.. TOTOBJ-sum(j,OBJ(j))=e=0;
QKONST(j,k).. sum(i,W(i,j)*Y(i,k))=l=1;
QWR(i,j,l).. (10*W(i,j)-W(l,j))*Y(i,j)*Y(l,j)=g=0;

MODEL MISD /all/

SOLVE MISD using LP Maximizing TOTOBJ;

DISPLAY TOTOBJ.l, OBJ.l, W.l;
```