V Congreso de Ingeniería de Organización
Valladolid-Burgos, 4-5 Septiembre 2003

Índice

Selección de inputs y outputs en el análisis envolvente de datos: una metodología basada en el análisis multivariante.

Carlos Serrano Cinca[1], Cecilio Mar Molinero[2]
[1] Profesor titular, Departamento de Contabilidad y Finanzas
Facultad de Económicas, Universidad de Zaragoza
serrano@posta.unizar.es
[2]Cecilio Mar Molinero
Becario Ramón y Cajal, Institut d'Organització i Control de Sistemes Industrials
Universitat Politecnica de Catalunya
Cecilio.mar@upc.es

RESUMEN

La selección de los inputs y outputs que entran en un modelo de Análisis Envolvente de Datos (DEA) es problemática. La eficiencia comparativa de cualquier unidad de decisión (DMU) depende de los inputs y outputs que entren en el modelo. Existe la tentación de hacer una reducción de datos basada en el estudio de las correlaciones. Pero la presencia o ausencia de variables altamente correlacionadas puede afectar las eficiencias calculadas de muchos DMUs. Por otra parte, la presencia de inputs u outputs que no se corresponden a la estructura lógica del sistema que se estudia puede hacer que algunos DMUs aparezcan como casos extremos en aspectos irrelevantes y parezcan ser eficientes por esa misma razón. Se han propuesto varios métodos para la selección de inputs y outputs, pero algunos no tienen un soporte teórico, mientras que otros dejan claro que dos modelos no difieren mucho en la estructura de las eficiencias calculadas pero no explican en qué modo son parecidos o diferentes. En este trabajo se presenta una metodología basada en la estimación de una multiplicidad de modelos, lo que genera una tabla de eficiencias por modelo y DMU. Esta tabla se estudia por medio del Análisis de Componentes Principales y otros métodos multivariantes que resultan en representaciones gráficas de los resultados. La metodología propuesta tiene varias ventajas: guía la selección de modelos en DEA, explica en qué modo dos modelos se parecen o se diferencian, explica porqué un mismo DMU adquiere distintas eficiencias bajo distintos modelos, visualiza los resultados, y produce un ordenamiento de DMUs incluso cuando los DMUs tienen 100% de eficiencia.

***Palabras clave**: Análisis Envolvente de Datos, DEA, Estadística Multivariante, Selección de modelos en DEA.*

Introduction.

Most researchers decide "a priori" what the specification of a DEA model should be, without considering any alternatives. But it is possible that a variable included may contribute little or nothing to the calculation of efficiency values. The converse is also true: a variable for which data is available, and has not been included in the model on a priori considerations, may be important in the determination of efficiencies. A methodology aimed at guiding model selection in DEA is clearly desirable. Two interesting model selection approaches are due to Norman and Stocker [1] and to Pastor et al [2]. Norman and Stocker (1991) assess the need to include a variable by correlating the values of the variable under consideration with efficiency values obtained from the model that excludes it. Pastor et al (2001) prove that the contribution of a variable to efficiency can be assessed by estimating efficiencies twice, once with the

reduced model -which does not include the variable-, and once with the total model -which includes the variable. However, as any empirical study demonstrates, models that appear to be similar are not exactly equivalent. As variables enter or leave the specification, some DMUs become 100% efficient or loose this characteristic. Both methodologies rely, to a certain extent, on judgement for final model selection. This judgement is made with little reference to the original data set, which becomes obscured in a mass of mathematical details. DEA provides, for each DMU, just a score. It is not very informative as to the way in which inputs and outputs contribute to the efficiency calculation. There are many ways of achieving similar levels of efficiency when various inputs and outputs are involved. It is necessary to look beyond DEA, study the reasons why DMUs achieve a certain degree of efficiency, and the reasons why the various models are, or are not, equivalent. Here we propose a methodology based on multivariate statistical analysis.

This paper proposes a new approach to guide model selection in DEA and to the ranking of units. The method has the advantage that the ranking extends to inefficient units. The DEA modelling procedure is embedded in a multivariate statistical framework. The procedure proposed attempts to visualise differences and similarities between the efficiencies generated by various DEA models. The model is developed within the context of the data set on Chinese cities studied by Zhu [3] and Premachandra [4].

2. Case study: Chinese cities and DEA.

Zhu's data set published on 18 Chinese was published by Premachandra. There are two inputs and three outputs defined as follows.

Input 1, (X1): Investment in fixed assets by state-owned enterprises.
Input 2, (X2): Foreign funds actually used
Output 1, (Y1): Total industrial output value
Output 2, (Y2): Total value of retail sales
Output 3, (Y3): Handling capacity of coastal ports

The first step in the procedure we propose here requires the listing of all possible DEA models that can be derived from possible inputs and outputs. These are shown in Table 1. To make it easy for identification purposes, notation is written in such a way that the inputs and outputs can be easily identified. In this way, the first input, X1, is associated with the letter A in the name; the second input, X2, is associated with the letter B; outputs are associated with numbers in an obvious way. Thus, model A1 in Table 2 contains one input, X1, and one output, Y1. Model A12 contains input X1 and outputs Y1 and Y2. Both Zhu and Premachandra only estimate AB123.

Efficiencies from each model were obtained using the CCR model with input orientation. Table 2 shows the efficiencies obtained. The influence of the model on efficiency can be clearly observed in Table 2. For example, DMU 2 is 100% efficient in twelve models that include output Y3 in their specification (A123, A13, A23, A3, B123, B13, B23, B3, AB123, AB13, AB23 and AB3). But if Y3 is removed from the specification, the efficiency of DMU 2 drops to very low values ranging from 0.11 to 0.33. Something similar could be said about

DMU 6 and DMU 10. Sometimes they are 100% efficient and other times they appear to be inefficient. In this case, as in every other, it is possible to scan through Table 2 in search of clues that may explain which inputs or outputs are responsible for the changes. It is, however, desirable to analyse Table 2 in a multivariate analysis context. Models can be treated as variables and efficiencies as observations. The aim is to explore the structure of the data and to visualise its most important features.

DMU	*INPUT*	*OUTPUT*
A1	X1	Y1
A12	X1	Y1 Y2
A123	X1	Y1 Y2 Y3
A13	X1	Y1 Y3
A23	X1	Y2 Y3
A2	X1	Y2
A3	X1	Y3
B1	X2	Y1
B12	X2	Y1 Y2
B123	X2	Y1 Y2 Y3
B13	X2	Y1 Y3
B23	X2	Y2 Y3
B2	X2	Y2
B3	X2	Y3
AB1	X1 X2	Y1
AB12	X1 X2	Y1 Y2
AB123	X1 X2	Y1 Y2 Y3
AB13	X1 X2	Y1 Y3
AB23	X1 X2	Y2 Y3
AB2	X1 X2	Y2
AB3	X1 X2	Y3

Table 1: The 21 DEA models used in the study.

It is clear that DMUs 2, 6, 10 are very different even if they all appear to be 100% efficient under the complete model, AB123. Another interesting example is provided by DMUs 1 and 16. A cursory examination of Table 2 suggests that they are not very different, and under the complete model AB123 they both achieve 47% efficiency. Are they similar? If they are not, where are the differences?

3. DEA and PCA. Efficiencies as variables in a multivariate statistical framework.

Models in Table 2 have been treated as variables and efficiency ratings as observations, and a PCA exercise has been performed. The minimum value for eigenvalue extraction has been set to 0.8, in line with Joliffe's [5] recommendation. Three eigenvalues exceeded the 0.8 limit, indicating that three components are sufficient to describe the structure of the data. The first component was by far the most important, accounting for 71.9% of the variability in the data.

The addition of the second component increases this percentage to 91.9%, and the addition of the third one takes it to 96.5%. This dominance of the first component is typical of highly correlated variables; [6]. For the purposes of this study, the first two components provide an adequate representation of the data. The results are given in Table 3.

DMU	*A1*	*A12*	*A123*	*A13*	*A23*	*A2*	*A3*	*B1*	*B12*	*B123*	*B13*	*B23*	*B2*	*B3*	*AB1*	*AB12*	*AB123*	*AB13*	*AB23*	*AB2*	*AB3*
1	27	28	47	47	44	25	26	10	10	12	12	8	6	3	27	28	47	47	44	25	26
2	11	17	100	100	100	17	100	33	33	100	100	100	32	100	33	33	100	100	100	32	100
3	26	26	28	28	21	19	5	9	9	9	9	4	4	1	26	26	28	28	21	19	5
4	37	37	50	50	41	28	20	17	17	19	19	10	8	3	37	37	50	50	41	28	20
5	48	54	63	59	58	49	17	52	52	53	53	34	32	6	59	59	63	63	58	49	17
6	100	100	100	100	88	88	13	84	84	84	84	44	44	3	100	100	100	100	88	88	13
7	28	28	36	36	26	18	12	9	9	10	10	4	4	1	28	28	36	36	26	18	12
8	22	33	50	41	49	32	25	16	16	19	19	17	14	6	22	33	50	41	49	32	25
9	50	50	63	63	48	35	21	53	53	55	55	26	22	7	60	60	66	66	48	35	21
10	60	100	100	66	100	100	14	100	100	100	100	100	100	8	100	100	100	100	100	100	14
11	20	24	30	27	28	22	11	9	9	10	10	7	6	2	20	24	30	27	28	22	11
12	26	59	79	52	79	59	32	12	16	19	15	19	16	5	26	59	79	52	79	59	32
13	32	70	75	43	75	70	16	21	27	28	22	28	27	3	32	70	75	43	75	70	16
14	12	13	14	14	12	11	3	3	3	3	3	2	1	0	12	13	14	14	12	11	3
15	15	18	19	16	17	17	3	2	2	2	2	1	1	0	15	18	19	16	17	17	3
16	26	46	47	30	47	46	7	4	5	5	4	5	5	0	26	46	47	30	47	46	7
17	27	27	31	31	27	23	8	2	2	2	2	1	1	0	27	27	31	31	27	23	8
18	5	17	20	10	20	17	6	2	4	5	3	5	4	1	5	17	20	10	20	17	6

Table 2: DEA efficiencies for DMU under the 21 DEA models. Efficiencies vary between 0 and 100.

Component	Eigenvalue	% of variance	Cumulative %
PC1	15.10	71.88	71.88
PC2	4.21	20.05	91.93
PC3	.95	4.51	96.45
PC4	.61	2.91	99.37

Table 3: PCA results.

Models have been ranked according to their loading in the first principal component. The results are shown in Table 4. All the models have positive loadings in this component. The model with the highest loading in the first component is AB123, the complete model that includes two inputs and three outputs. In this kind of situation the first principal component is often taken to be an overall measure of strength of the relationship. It is clear that this component can be interpreted as an overall measure of efficiency. Ranking of DMUs on this component will produce a ranking of all DMUs in terms of efficiency; this ranking includes both efficient and inefficient DMUs

	Component		
	PC1	PC2	PC3
AB123	.964	.141	.197
A123	.963	.143	.205
AB13	.962	.173	-.122
B123	.942	.204	-.246
AB23	.937	.187	.290
A23	.937	.187	.290
B13	.932	.207	-.284
B12	.923	-.263	-.256
B1	.913	-.256	-.303
B2	.894	-.170	-.100
AB12	.893	-.411	.146
A13	.891	.281	-
AB1	.882	-.366	-.266
B23	.874	.345	-.116
AB2	.872	-.374	.288
A12	.827	-.525	.195
A2	.815	-.487	.293
A1	.730	-.530	-.187
B3	.397	.896	-
AB3	.438	.890	-
A3	.438	.890	-

Table 4: Component loadings. Models are ordered on the first component.

Turning to the second component, it is to be noticed that the only models that load highly on it are B3, AB3 and A3. All these models contain a single output in their specification, Y3. All the models that contain output Y3 have positive loadings in the second component, while those models that exclude Y3 have negative loadings. The second component is clearly associated with the ability that DMUs have of generating output Y3. Using similar reasoning, it can be argued that the third principal component is related to the efficient use of inputs. Models A2, A23, AB23, AB2, A123, AB123, A12, AB12, A3, and AB3 have positive loadings in the third components. All these models contain input X1 or both inputs. Models B1, B13, AB1, B12, B123, A1, AB13, B23, B2, B3, and A13 have negative loadings in the third component. All but two models contain input X2 or both inputs.

In summary, the first principal component gives an overall measure of efficiency; the second principal component is related to output Y3; and the third principal component is a contrast between input X1 and input X2. Here we will concentrate on the first two components.

For each DMU, component scores for the first and second principal component have been plotted in Figure 1. The DMUs that achieve efficiency scores of 100% are to be found at the extreme right hand side of the first principal component. DMU 2 shows its reliance on output Y2 by finding its way to the top of the second principal component. The fact that DMU 2 achieves efficiency by concentrating on output Y3 is now clear. The isolated position of DMU 2 in the figure suggests that we are dealing with a "maverick". In general, once meaning has been attached to the various components, extreme points can be analysed, particularly efficient extreme points, as this may indicate that the relevant DMUs use an unusual mix of inputs and outputs to achieve efficiency, and this may reveal maverick behaviour. At the other extreme of the first principal component we find DMUs 14, 15, and 18. These DMUs achieve low efficiencies under most models.

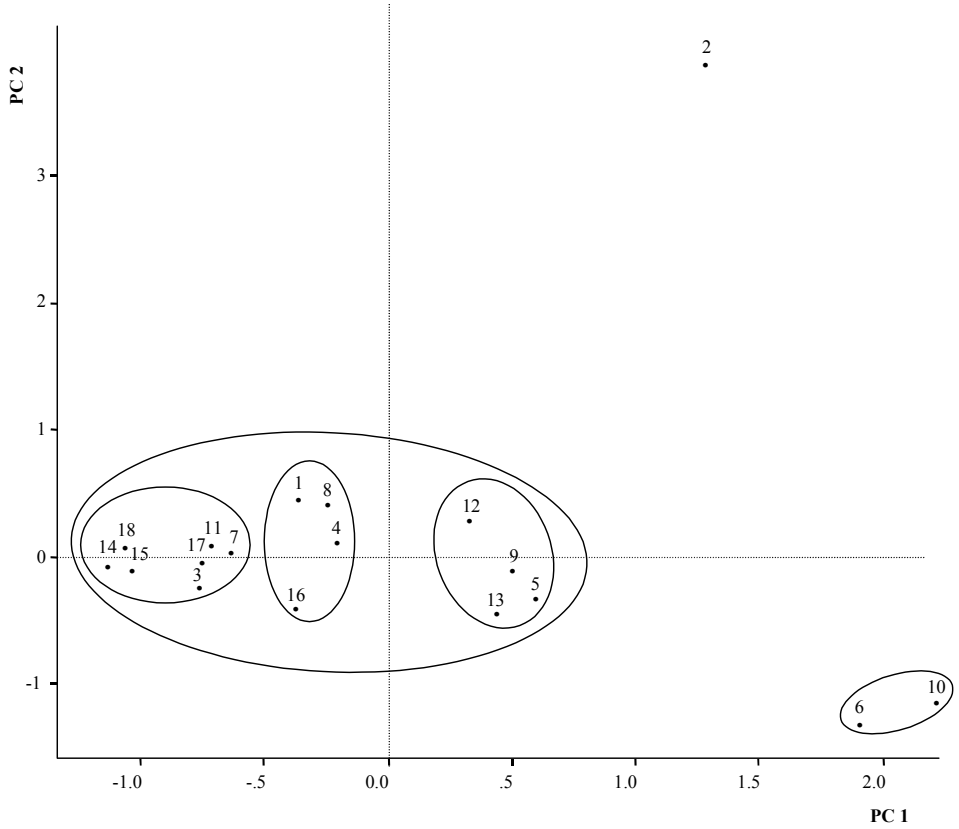

Figure 1. Component scores for the first and second principal component

The relationship between components and models can be displayed graphically by using the technique of Property Fitting (Pro-Fit); [7]. In this technique, vectors are drawn in such a way that, for a particular DEA model, the value of the efficiency derived from the model increases in the direction of the vector. The direction of the vector is calculated as a result of a regression analysis in which the efficiencies derived from the particular model are the dependent variables and the component scores are the independent variables. This technique has the advantage of highlighting up to what point two models are similar, since the angle between any two vectors is related to the correlation between the efficiencies generated by the two models concerned. All the vectors are represented through the centre of coordinates in Figure 1. In this case all models achieved very high values of R^2, the lowest one being 0.81, and all models were represented. All vectors pointed towards the positive side of the first component, forming an open fan, something that indicates that the various ways of achieving efficiency are positively correlated. The vectors can be seen in Figure 2.

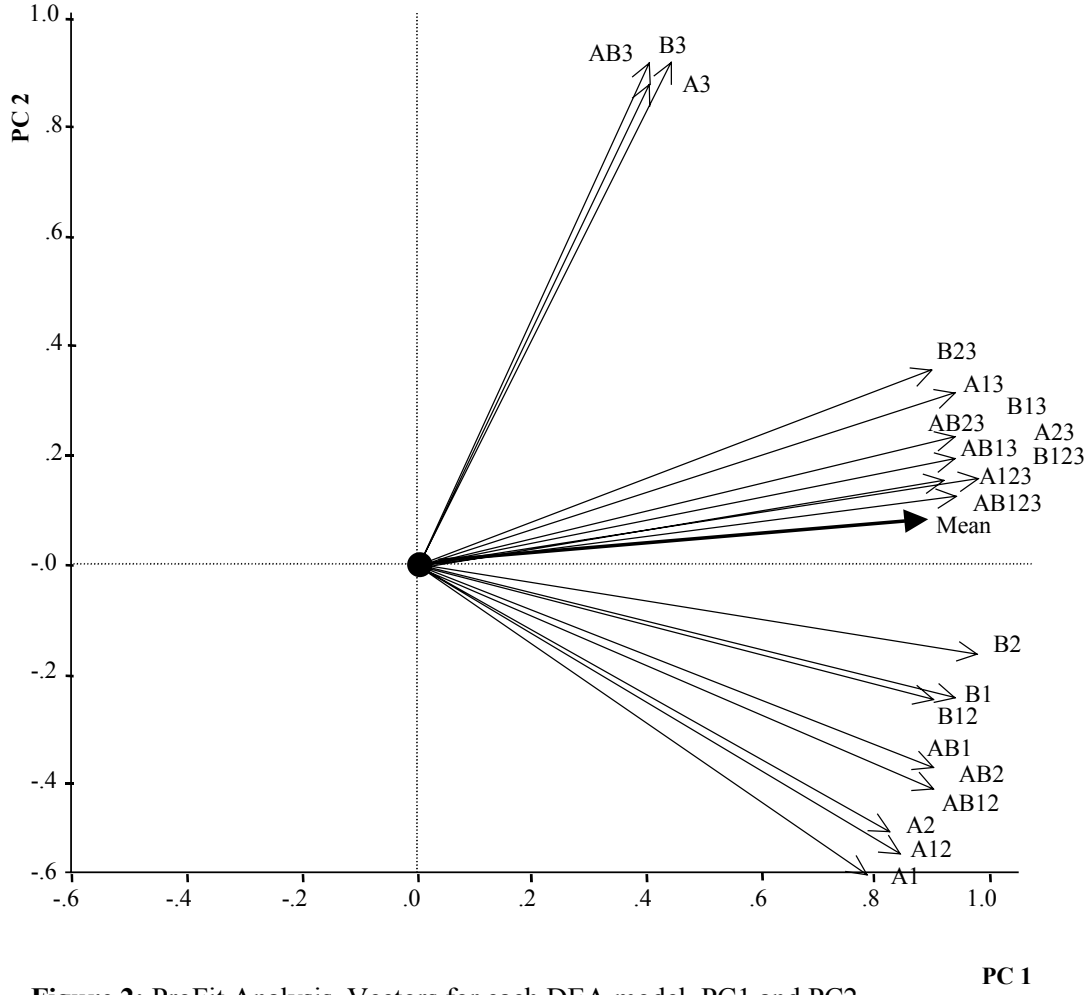

Figure 2: ProFit Analysis. Vectors for each DEA model. PC1 and PC2.

The vectors in Figure 2 group neatly into three groups. One group is formed by models AB3, B3, and A3. All these models achieve their highest value in DMU 2 and contain only output 3, indicating that DMU 2 achieves 100% efficiency by attaching high weights to output 3. The remaining vectors split into two groups, both of them pointing in the positive direction of

component 1. One group points towards the positive side of component 2 and the other group points towards the negative side of component 2. The difference between the two groups concerns the presence or absence or output 3 in the specification. The models that do not contain output 3 point downwards, and those that contain output 3 point upwards. Thus, output 3 is crucial in the modelling procedure. The average vector -labelled Mean- has also been calculated and represented, and almost coincides with the axis associated with the first component. In other practical situations one would also look at the projection on other principal components, and this may reveal the different reasons why DMUs achieve a given level of efficiency.

The procedure to select a model is now clear. If the directional vectors form a closed fan, model selection is very simple, as this is an indication that all models are equivalent. In this case one would select the most parsimonious model. If the fan is wide open, we need to explore any groups that may exist and base our model selection on economic considerations as well as on statistical principles. Thus, the fan is the wind rose that guides the DEA sailor through the sea of models. In the present case it is to be first decided whether output 3 should or should not be included in the specification. This is a crucial decision. Models AB3, B3, and A3 do not appear to be reasonable since they favour a maverick DMU, DMU 2, and show the remaining DMUs in bad light, a fact that can be confirmed by inspecting Table 3. If it is decided to leave output 3, in the specification, then any model amongst B23, A13, B13, AB23, B123, A23, B123, A123, AB123 could be chosen. Parsimony would probably favour A23, as it plots in the middle of the fan and, contains only one input and two outputs.

We can now see in which way DMU 1 is different from DMU 16. They both achieve the same efficiency score under the complete model AB123, and have almost identical projections on the first principal component. But DMU 1 plots on the positive side of the second component, indicating that it values output Y3, while DMU 16 plots towards the negative side of the second principal component, indicating that models that ignore output Y3 will favour this DMU. If output Y3 was to be considered important by decision makers, DMU 1 would be preferred to DMU 16.

As far as DMU ranking is concerned, it could be argued that no single model should contribute to the position of a DMU in the list, and that the ranking should take into account all possible specifications. Thus, the ranking along the first principal component would be appropriate. We think that only the first principal component should be involved in the ranking, and not all of them weighted according to the variance they explain, as done by Zhu. The ranking based on the first principal component would produce the following ordering of DMUs: 10, 6, 2, 5, 9, 13, 12, 4, 8, 1, 16, 7, 11, 17, 3, 15, 18, and 14. It is to be noticed that this procedure allows for the ranking of all DMUs: inefficient and efficient ones.

A complementary way of analysing the data in Table 3 is to use Cluster Analysis. It is good practice to supplement the results obtained from graphical representations of multivariate data with the superimposition of cluster countours; [8]. Clusters were obtained using Ward's method, that maximises within group homogeneity and between group heterogeneity. Cluster Analysis shows the presence of five main clusters, one of them containing only DMU 2, which appears yet again as a special case. At a higher level of clustering, DMUs divide neatly into two groups, one of them containing DMUs that reach 100% efficiency and the other one containing the DMUs that never do. The clusters are shown in Figure 1.

The extreme position of DMUs 14, 15, and 18 is to be noted. Zhu comments on these three cities as follows: "these three DMUs were declared by Chinese government as model for economic reforms and developments". Considering the low efficiency levels achieved by these three cities, any directed economic effort has great opportunities for success.

4 Conclusion.

This paper has presented a new method for model selection in DEA based on multivariate statistical analysis. The methodology requires evaluating efficiencies for all possible input/output combinations. It is clear that such methodology produces much redundancy, but also generates valuable information. The matrix of efficiencies by models is then analysed by means of data reduction techniques, such as Principal Components Analysis. Further understanding of the data can be gained by applying Hierarchical Cluster Analysis in this data set.

It has been shown that there are advantages with calculating efficiencies under all possible specifications of the DEA model, and then performing multivariate analysis on the results obtained. This methodology permits the joint graphical representation of models and DMUs, and thus it makes it possible to explain up to what point two models are equivalent, and if they are not equivalent, why they are not equivalent. The relationship between models and DMUs becomes clarified. By supplementing the representations with the results of Property Fitting techniques, it is possible to assess why a particular DMU achieves high efficiency scores under some models and low efficiency scores under other models. Maverick DMUs are easily identified. Finally, the method permits the ranking of DMUs. Such ranking includes both efficient and inefficient DMUs.

Agradecimientos

La participación de Cecilio Mar en este trabajo ha sido posible gracias a una beca Ramón y Cajal del Ministerio de Ciencia y Tecnología.

Referencias

[1] Norman, M. and Stocker, B. (1991): *Data Envelopment Analysis: the assessment of performance.* John Wiley and Sons. Chichester. UK.

[2] Pastor, J.; Ruiz Gomez, J.L; and Sirvent, I. (2001): A Statistical Test for Nested Radial DEA Models. Operations Research, 50, 728-35.

[3] Zhu, Joe (1998): Data envelopment analysis vs. principal component analysis: An illustrative study of economic performance of Chinese cities. European Journal of Ope*rational Research*, (111): 1, pp: 50-61.

[4] Premachandra, I M (2001): A note on DEA vs principal component analysis: An improvement to Joe Zhu's approach, *European Journal of Operational Research*, (132): 3, pp: 553-560.

[5] Joliffe. I.T. (1972): Discarding variables in Principal Components Analysis. *Applied Statistics*, 21, 160-173.

[6] Dunteman G.H. (1989): *Principal Component Analysis*. Sage Publications Ltd. London, UK.

[7] Schiffman, J.F., Reynolds, M.L. and Young, F.W. (1981): *Introduction to Multidimensional Scaling: Theory, Methods and Applications*. Academic Press, London.

[8] Arabie P., Carroll J.D. and De Sarbo W.S. (1987): *Three way scaling and clustering*. Sage University Paper Series on Quantitative Applications in the Social Sciences. Number 07-065. Beverley Hills. Sage Publications.