Combined Model for Clustering Observations of Systems Subject to Inefficiency

02991984 Osaka University *Ilyes BELLAMINE
01604524 Osaka University Hiroshi MORITA
01005194 Osaka University Hiroaki ISHII

Introduction:

This paper is concerned with the regrouping of clusters of systems subject to inefficiency into groups whose observations share common characteristics such as the performance level. As defined in [4, 5] the inefficiency is a systematic term due to change of some intrinsic properties of the input such as the change of some physical characteristics of a mix component and/or a change of settings of some supposed fairly stable (or mastered) factors whose are firstly confused into the disturbance term. Examples of inefficiency are detailed in [2, 3]. In previous works, having been inspired from the stochastic frontier model, we proposed a parametric technique to estimate the inefficiency function for linear systems [2, 3, 4, 5]. In those studies the systematic term, which represents the mix of input has been decomposed into two terms one representing the best output *obtained* from the mix of the used input and the other a positive term representing a lack of output caused by the so called inefficiency. For linear systems, the likelihood function has been maximized and the number of gaps of the inefficiency function has been optimized through AIC (Akaike Information Criterion) score.

The estimation of number of gaps returns to grouping observations into clusters, which will hopefully share common characteristics. The grouping methods performed on a raw data subject to inefficiency showed its limit. In this study, without assuming the linearity of the system, we propose a new non-parametric technique, which combine BCC (Banker-Charnes-Cooper) model and the clustering methods. We will show in the following that clustering of BCC-scores of the observations is more meaningful than clustering of the raw data. In addition, from the scores dendrograms, the number and positions of gaps are more easily found than by the parametric method. Good results have been found for a data subject to a single and multiple step inefficiency functions. Following this introduction, the BCC-model is introduced and the new concept is developed.

The BCC model

BCC model is one of the most basic DEA (Data Envelopment Analysis) models [1], which was initially proposed by Banker, Charnes and Rhodes [6]. The production frontiers of the BCC-model are spanned by the convex hull of the existing DMUs (Decision making Units) and have piece-wise linear characteristics. The production possibility set in the area consisting of the frontier together with observed activities with an excess of input and/or a shortfall in output compared with the frontiers. In figure 1, the frontiers of BCC-model consists of the bold lines connecting A, B and C. The BCC-efficiency score of D is evaluated by the ratio PR/PD. A, B and C are on the frontiers, so they are BCC efficient and their BBC-efficiency scores are 1.

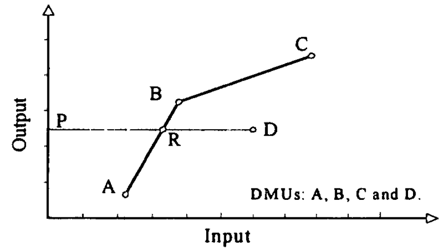

Figure 1: The BCC model.

The combined model

Let us begin with a simple example. Figure 2 exhibits a system subject to inefficiency characterized by a shortage of output for the observations 7~20. For this simple case, "data clustering" would be the best method to regroup the data into three groups. However, from the dendrogram plotted in figure 3 this task is not very easily accomplishable. Because the similarity level between the observations is very high, which means that the distance between clusters is very small. Therefore the final regrouping of the observations into three clusters cannot be easily done.

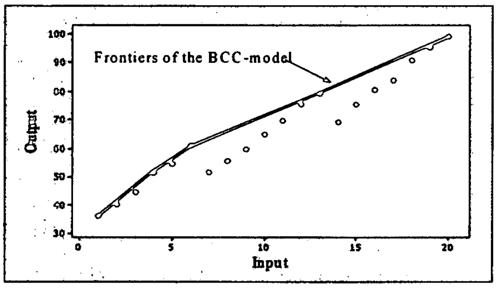

Figure 2: Scatter plot of the raw data.

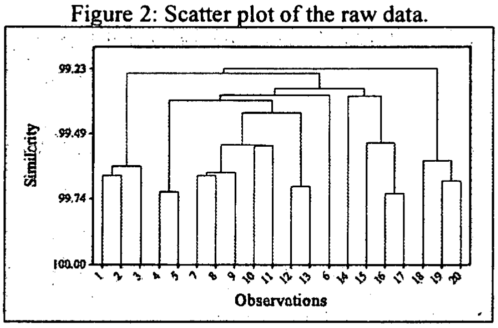

Figure 3: Dendrogram of the raw data of one-independent variable system for the similarity level 95%.

We remember that our purpose is to estimate the relative inefficiency. Then, we have to compare the performance of the observations between each other and group the data into clusters having nearly the same performance level. For this purpose, we propose to estimate the BCC-efficiency scores of the different observations and cluster them rather than clustering directly the observations of the raw data.

It is to remark that except the extremities, the BCC-efficient observations are points of change of performance, and, we may think that these points could be on the edge of clusters whose observations share common level of performance. Therefore, to determine the similarity level for cutting the dendrogram, the BCC-efficient observations are the key points, which can help us to make the most sense to our data.

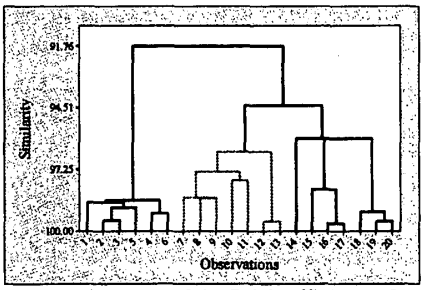

Figure 4: Dendrogram of the BCC-efficiency Scores of one-independent variable system for the similarity level 95%.

Using this proposed technique, which combine the BCC-model with the clustering method, we were able to distinguish more easily the three groups whose observations are characterized by nearly the same performance level. As shown in Figure 4, compared to the dendrogram of the raw data the minimum level of similarity has been improved from 99.23 to 91.76%, which means that the distance between the observations increased, consequently, the clustering process became easier. For the same level of similarity set to 95%, the clustering of the raw data classified the observations into only one group; however, the clustering of the BCC-efficiency scores classified the observations into exactly the expected three groups.

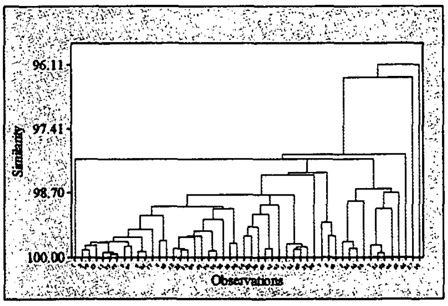

Figure 5: Dendrogram of the raw data of a two-independent variable system for the similarity level 95%.

Figure 5 and 6 illustrates the dendrograms of a two-independent variable system, where only an independent variable is subject to inefficiency. The dendrograms of the BCC-scores shows as expected the presence of two clusters, characterized by the same level of performance.

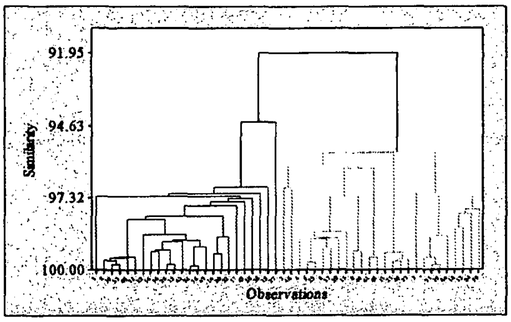

Figure 6: Dendrogram of the BCC-efficiency Scores of a two-independent variable system for the similarity level 95%

Conclusion

A combined model has been described for clustering observations of systems subject to inefficiency. The main advantage of the proposed model is the grouping of observations into clusters, which share nearly the same performance level. For systems described by one and two independent variables it was shown that clustering of the BCC-efficiency scores is more meaningful and easier to interpret than the clustering of the raw data. Future research will focus more on clustering of observations of multi-variable independent systems.

This work has been partly supported by the Sasagawa Scientific Research Grant from The Japan Science Society.

References

[1] Bauer PW. (1990). Recent Development in the Econometric Estimation of Frontiers, Journal of Econometrics, 46, pp. 39-56.

[2] Bellamine I, Morita H. and Ishii H. (2003). Concept of the Systematic Inefficiency and its Applications, Proceedings of the 2003 Fall National Conference of the Operations Research Society of Japan, pp. 254-255.

[3] Bellamine I and Morita H. (2003). The Concept of Inefficiency in Quality Control for the Continual Improvement, Proceedings of the 73rd Research Presentation Meeting in Kansai, The Japanese Society for Quality Control, pp. 1-4.

[4] Bellamine I, Morita H. and Ishii H. (2003). Inefficiency and Model Simplification in Quality Control, Proceedings of the Fifth International Industrial Engineering Conference - Quebec, ISBN 2-9808240-0-3.

[5] Bellamine I, Morita H. and Ishii H. (2003). An Estimation Method of the Parameters of a Linear Regression System Subject to Inefficiency, Proc. the 9th APMC, pp. 567-577

[6] Banker R. D., Charnes A. and Cooper W.W. (1984). Some Models for Estimating Technical and Scale Inefficiencies in Data Envelopment Analysis. Management science, 30, pp. 1078-1092.