Chapter 17

PREPARING YOUR DATA FOR DEA

Joe Sarkis
Graduate School of Management, Clark University, 950 Main Street, Worcester, MA, 01610-1477, jsarkis@clarku.edu

Abstract: DEA and its appropriate applications are heavily dependent on the data set that is used as an input to the productivity model. As we now know there are numerous models based on DEA. However, there are certain characteristics of data that may not be acceptable for the execution of DEA models. In this chapter we shall look at some data requirements and characteristics that may ease the execution of the models and the interpretation of results. The lessons and ideas presented here are based on a number of experiences and considerations for DEA. We shall not get into the appropriate selection and development of models, such as what is used for input or output data, but focus more on the type of data and the numerical characteristics of this data.

Key words: Data Envelopment Analysis (DEA), Homogeneity, Negative, Discretionary

1. SELECTION OF INPUTS AND OUTPUTS AND NUMBER OF DMUS

Selection of inputs and outputs and number of DMUs is one of the core difficulties in developing a productivity model and in preparation of the data. In this brief review, we will not focus on the managerial reasoning for selection of input and output factors, but more on the computational and data aspects of this selection process.

Typically, the choice and the number of inputs and outputs, and the DMUs determine how good of a discrimination exists between efficient and inefficient units. There are two conflicting considerations when evaluating the size of the data set. One consideration is to include as many DMUs as
possible because with a larger population there is a greater probability of capturing high performance units that would determine the efficient frontier and improve discriminatory power. The other conflicting consideration with a large data set is that the homogeneity of the data set may decrease, meaning that some exogenous impacts of no interest to the analyst or beyond control of the manager may affect the results (Golany and Roll 1989; Haas and Murphy, 2003). Also, the computational requirements would tend to increase with larger data sets. Yet, there are some rules of thumb on the number of inputs and outputs to select and their relation to the number of DMUs.

**Homogeneity**

There are methods to look into homogeneity based on pre-processing analysis of the statistical distribution of data sets and removing “outliers” or clustering analysis, and post-processing analysis such as multi-tiered DEA (Barr, et al., 1994) and returns-to-scale analysis to determine if homogeneity of data sets is lacking. However, multi-tiered approaches require large numbers of DMU to do this.

Another set of three strategies to adjust for non-homogeneity were proposed by Haas and Murphy (2003). The first, a multi-stage approach by Sexton et al. (1994) is one technique. In the first stage they perform DEA using raw data producing a set of efficiency scores for all DMUs. In the second stage they run a stepwise multiple regression on that set of efficiency scores using a set of site (exogenous) characteristics that are expected to account for differences in efficiency. In the third stage they adjust DMU outputs to account for the differences in site characteristics and perform a second DEA to produce a new set of efficiency scores based on the adjusted data. The adjusted output levels used in the second DEA are derived by multiplying the level of each output by the ratio of the DMU’s unadjusted efficiency score to its expected efficiency score. The magnitude of error method (actual minus forecast) is the second technique suggested by Haas and Murphy (2003). In this approach the adjustment is to estimate inputs and outputs using regression analysis. The factors causing non-homogeneity are treated as independent variable(s). DEA is then applied using the differences between actual and forecast inputs and outputs, rather than the original inputs and outputs. The ratio of actual to forecast method (actual/forecast) is the third technique suggested by Haas and Murphy (2003). This time instead of using differences, they use the ratio of actual input to forecast input and actual outputs to forecast outputs executing DEA. But simulation experiments showed that none of these three strategies performed well. At this time alternative approaches for addressing non-homogeneity are limited.