Chapter 16

DATA ENVELOPMENT ANALYSIS WITH MISSING DATA
A Reliable Solution Method

Chiang Kao¹ and Shiang-Tai Liu²
¹Department of Industrial and Information Management, National Cheng Kung University, Tainan 701, Taiwan, Republic of China, ckao@mail.ncku.edu.tw
²Graduate School of Business and Management, Vamung University, Chung-Li, Tao-Yuan 320, Taiwan, Republic of China, stliu@vnu.edu.tw

Abstract: In data envelopment analysis (DEA), the input and output data from all of the decision making units (DMUs) to be compared are required. If, for any reason, some data are missing, then the associated DMU must be eliminated to make the approach applicable. This study proposes a fuzzy set approach to deal with missing values. The value of a DMU in an input (or output) which is missing is represented by a triangular fuzzy number constructed from the values of other DMUs in that input (or output). A fuzzy DEA model is then used to calculate the efficiencies, which are usually also fuzzy numbers. We use a problem with complete data to investigate the effect of this approach when 1%, 2%, and 5% of the values are missing. While the conventional DMU-deletion method will overestimate the efficiencies of the remaining DMUs, the fuzzy set approach produces results which are very close to those calculated from complete data. The average error in estimating the true efficiency is less than 0.3%. Most importantly, the fuzzy set approach is able to calculate the efficiencies of all DMUs, including those with some values missing.

Key words: Data envelopment analysis, efficiency, missing data, fuzzy set

1. INTRODUCTION

Since the pioneering work of Charnes et al. (1978), data envelopment analysis (DEA) has been widely studied from both the theoretical and
practical points of view. Different models have been developed to measure the efficiency of a group of decision making units (DMUs) which utilize the same inputs to produce the same outputs under different conditions. Applications for different types of organizations have also been reported (Seiford 1996, 1997, Cooper et al. 2000).

The basic idea of DEA is to allow each decision making unit to use different virtual multipliers, the most favorable, in calculating its relative efficiency expressed as the ratio of aggregated output to aggregated input. In selecting the multipliers, it is required that the ratio of aggregated output to aggregated input calculated from the multipliers selected by the DMU concerned should not exceed 1.0 for all DMUs. Let $X_{ij}, j=1,\ldots, s$ and $Y_{ik}, k=1,\ldots, t$ denote the $j$th input and $k$th output, respectively, of DMU $i, i=1,\ldots, n$. Banker et al. (1984) develop the following mathematical program to calculate the efficiency of DMU $r$:

\[
E_r = \max \sum_{k=1}^{t} u_k Y_{rk} / (v_0 + \sum_{j=1}^{s} v_j X_{ij})
\]

\[
\text{s.t. } \sum_{k=1}^{t} u_k Y_{ik} / (v_0 + \sum_{j=1}^{s} v_j X_{ij}) \leq 1, \quad i = 1,\ldots, n \tag{1}
\]

where $u_k$ and $v_j$ are the multipliers associated with output $k$ and input $j$, respectively, to be determined from (1) and $\varepsilon$ is a small non-Archimedean number (Charnes et al. 1979, Charnes and Cooper 1984) imposed to avoid DMU $r$ from assigning zero weight to unfavorable factors. $E_r$ is the relative efficiency of DMU $r$, where $E_r=1$ indicates efficiency and $E_r<1$ inefficiency. Model (1) is a linear fractional program which can be solved by transforming to a linear program (Charnes and Cooper 1962) and utilizing any linear programming solver. The underlying assumption of this model is variable returns to scale. If $v_0$ is set to zero, then the model boils down to one under the assumption of constant returns to scale.

As indicated by its name, data envelopment analysis is based on data. To calculate the relative efficiency, data from all DMUs are required. If any observation of a DMU in the group is missing, then this DMU must be deleted from the group in order to calculate the efficiency of all other DMUs. A consequence of this is overestimation of the efficiency of some DMUs, because the number of DMUs for comparison is decreased. As more DMUs are deleted due to lack of data, the resulting efficiencies will be biased high to a larger extent. Therefore, it is desirable to keep those DMUs with some observations missing in the group by making some amendments.

The problem of missing data has been widely discussed in statistical analysis (Allison 2002, Rubin 2004, Schafer 1997). There are formulas for