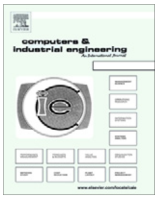
Contents lists available at SciVerse ScienceDirect

Computers & Industrial Engineering

journal homepage: www.elsevier.com/locate/caie



Data Envelopment Analysis of clinics with sparse data: Fuzzy clustering approach ☆

David Ben-Arieh *, Deep Kumar Gullipalli

Department of Industrial and Manufacturing Systems Engineering, Kansas State University, Manhattan, KS 66502, United States

ARTICLE INFO

Article history:
Received 7 March 2011
Received in revised form 15 December 2011
Accepted 16 January 2012
Available online 28 January 2012

Keywords:
Data Envelopment Analysis
Sparse data
Clustering
Fuzzy c-means
Healthcare

ABSTRACT

This paper presents a method for utilizing Data Envelopment Analysis (DEA) with sparse input and output data using fuzzy clustering concepts. DEA, a methodology to assess relative technical efficiency of production units is susceptible to missing data, thus, creating a need to supplement sparse data in a reliable and accurate manner. The approach presented is based on a modified fuzzy c-means clustering using optimal completion strategy (OCS) algorithm. This particular algorithm is sensitive to the initial values chosen to substitute missing values and also to the selected number of clusters. Therefore, this paper proposes an approach to estimate the missing values using the OCS algorithm, while considering the issue of initial values and cluster size. This approach is demonstrated on a real and complete dataset of 22 rural clinics in the State of Kansas, assuming varying levels of missing data. Results show the effect of the clustering based approach on the data recovered considering the amount and type of missing data. Moreover, the paper shows the effect that the recovered data has on the DEA scores.

1. Introduction

DEA is a linear programming model, which measures the relative technical efficiency of decision making units by calculating the ratio of weighted sum of its outputs to its inputs (Charnes, Cooper, & Rhodes, 1978). Decision making units (DMUs) can be defined as any production unit, in any for-profit or non-profit organizations, which consumes inputs and produces outputs. The DEA model is run n times by changing the objective function each time to determine the best set of weights which maximize the efficiency of the DMU under evaluation, while the weights should remain feasible for all the other DMUs. DEA not only measures efficiency but also the amount of inefficiencies associated with each DMU by comparing inefficient DMUs against efficient DMUs. By solving the DEA model one can also obtain projection scores which represent the required increase in output or decrease in input for a DMU to be fully efficient. DEA is widely recognized as an effective method for measuring the relative efficiency of DMUs using a set of multiple inputs and multiple outputs. Extension of this particular methodology and its application to vast number of fields since its inception is presented in the works of Seiford (1997) and Emrouznejad, Parker, and Tavares (2008).

The area of health care operations is very suitable for DEA analysis since clinics (or any health providing organization) are easily defined as DMUs in the DEA context. The DEA analysis can accurately show the efficient aspects of the clinics as well as areas that need improvements. This work is based on a DEA analysis of clinics in Kansas that serve the rural and medically underserved population. One of the early findings of this research was that due to a lack of reporting standards each clinic may collect or report a different set of data items. Thus, when conducting a DEA analysis, it is common to find that some data items are not collected or collected inappropriately, creating the issue of missing data.

The application of DEA analysis in health care started as one of the earliest application domain. Analysis performed on American institutions include analysis of hospitals in Wisconsin (Nunamaker, 1983), inefficiencies in clinics (Sherman, 1984), physician efficiency (Ozcan, 1998), neurotrauma patients in the ICU (Nathanson, Higgins, Giglio, Munshi, & Steingrub, 2003), health maintenance organizations (Siddharthan, Ahern, & Rosenman, 2000), operating room efficiency (Basson & Butler, 2006), and local health departments in US (Mukherjee, Santerre, & Zhang, 2010). DEA applications outside the US include efficiency of nursing homes in Italy (Garavaglia, Lettieri, Agasisti, & Lopez, 2011), measured productivity of hospitals in Holland (Blank & Valdmanis, 2010), efficiency of public hospitals in Thailand (Puenpatom & Rosenman, 2008), efficiency of hospitals in Austria and Germany (Helmig & Lapsley, 2001; Hofmarcher, Paterson, & Riedel, 2002), and efficiency of long term care nursing care units in Finland (Björkgren, Häkkinen, & Linna, 2001) are a few examples.

The research presented here was used primarily to evaluate the efficiency of 41 KAMU (Kansas Association for the Medically Underserved) clinics which include 19 federally supported clinics, 14 primary care clinics, seven free clinics, and one voucher program. KAMU provides advocacy as well as training, technical assistance, and communication services to the clinics in an attempt to

☆ This manuscript was processed by Area Editor Imed Kacem.

* Corresponding author. Tel.: +1 913 532 5606.
E-mail address: davidbe@ksu.edu (D. Ben-Arieh).

0360-8352/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cie.2012.01.009

develop best practices. The purpose of this DEA analysis was to identify benchmarks and provide budget and resource recommendations for inefficient clinics. The clinics used a data reporting tool that collected up to 225 attributes. However, we found that a large amount of data was sporadically missing since each clinic collected a different subset of the data. In this study we reduced the data analyzed to 13 parameters that deemed essential for the DEA study and then developed the methodology presented herein to replace the missing data.

This paper explores a solution approach towards generating the missing data based on fuzzy clustering. Moreover, the paper demonstrates the sensitivity of this approach to the initialization process and to the cluster sizes chosen. The paper then shows the effect of this approach on the data recovered as well as on the DEA results. This contribution can help researchers improve the accuracy of the DEA analysis by generating the missing values more accurately, and also by understanding the effect of this approach on the DEA scores.

This paper is structured as follows: Section 2 provides a background and literature review of DEA and clustering approaches. Section 3 presents approaches for clustering with missing data, and Section 4 presents experimental results on the effect of the initial values as well as cluster sizes on the accuracy of the data recovered. Section 5 demonstrates the data generation approach using the actual clinical data with various patterns of missing values. Section 6 shows the effect of the data recovery strategy on the DEA analysis. Section 7 provides summary and conclusions.

2. Background

This section presents an introduction to basic DEA models, literature review of existing methods to handle missing values in DEA, and as well as an introduction to clustering approaches and the basic clustering algorithms.

2.1. Introduction to DEA models

Common DEA notations:

DEA	Data Envelopment Analysis
DMU	Decision making unit, a unit which consume inputs and produce outputs
DMU_o	DMU under evaluation or test DMU
n	Total number of DMUs under evaluation
m	Total number of input variables
s	Total number of output variables
*	Optimal solution value
v_i	Input multiplier variable of ratio model, $\forall$ $i = 1, 2, \ldots, m$
u_r	Output multiplier variable of ratio model, $\forall$ $r = 1, 2, \ldots, s$
X	Matrix representation of input variables
Y	Matrix representation of output variables
x_{ji}	Represents input variables of DMU_j, $\forall\ i = 1, 2, \ldots, m$
y_{jr}	Represents output variables of DMU_j, $\forall\ i = 1, 2, \ldots, s$
$[X_j, Y_J]$	Vector of inputs and outputs for DMU_J
$[X_o, Y_o]$	Vector of inputs and outputs for DMU_o

Consider a dataset of n DMUs which consume m inputs and produce s outputs. Input and output data for DMU_j are represented as, $x_{ji}(i = 1, 2, \ldots, m)$, and $y_{jr}(i = 1, 2, \ldots, s)$ respectively, where $(i = 1, 2, \ldots, n)$. Efficiency of each DMU is evaluated relative to the constraint set of all n DMUs, and needs n optimizations. DMU under evaluation is represented by DMU_o input and output vectors are represented as $[X_oY_o]$. The values u_r, v_i represent output and input weights of the multiplier model respectively.

Table 1
Basic DEA formulations – multiplier approach.

CCR input oriented model	CCR output oriented model
$Max\quad Z = \sum_{r=1}^{s} u_r y_{or}$ $S.to\quad (1)$	$Min\quad Z = \sum_{i=1}^{m} v_i x_{oi}$ $S.to\quad (2)$
$\sum_{i=1}^{m} v_i x_{oi} = 1$ $-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \quad \leqslant 0 \forall j = 1, \ldots, n$ $u_r, v_i \geqslant \forall r = 1, \ldots, s,\ i = 1, \ldots, m$	$\sum_{r=1}^{s} u_r y_{or} = 1$ $-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \quad \leqslant 0$ $u_r, v_i \geqslant \forall r = 1, \ldots, s,\ i = 1, \ldots, m$

Charnes et al. (1978) developed the first model (known as CCR). This model can be classified into an input or output oriented model. Input oriented models aim at minimizing the inputs with no change of outputs, whereas output oriented models aim at maximizing the outputs with no increase of inputs (Cooper, Seiford, & Tone, 2000). CCR model is based on constant returns to scale (CRS). The basic formulations of CCR input and CCR output models are shown in Table 1.

Banker, Charnes, and Cooper (1984) modified the CCR model creating the BCC model which employs variable return to scale (VRS). It assumes that there exists a variable proportional change between inputs and outputs. The BCC model has the production frontier spanning the convex hull of the existing DMUs. This frontier has piecewise linear and concave characteristics leading to the variable return to scale characteristics.

This paper considers only the CCR input model (model 1) for analysis. There are several other models of DEA such as Multiplicative Model (Charnes, Cooper, Seiford, & Stutz, 1982), Additive Model (Charnes, Cooper, Golany, Seiford, & Stutz, 1985), Assurance Region Model (Thompson, Singleton, Thrall, Smith, & Wilson, 1986), Cone Ratio Envelopment Model (Charnes, Cooper, Wei, & Huang, 1989), Malmquist Index (Fare & Grosskopf, 1992), and Super Efficiency Model (Andersen & Petersen, 1993) among many others. Each such particular model has specific advantages when compared to the basic CCR model.

2.2. DEA with missing data

The classical assumption of DEA is availability of numerical data for each input and output, with the data assumed to be positive for all DMUs (Cooper et al., 2000). This particular assumption limits the applicability of the DEA methodology to real world problems which contain missing values either due to human errors or technical problems.

In order to allow DEA analysis with missing data, minimal data requirements were defined. These requirements state that at least one DMU should have a complete set of inputs and outputs and each DMU should have at least one input and one output (Fare & Grosskopf, 2002). The accuracy of the results depends on the quality and quantity of the data. The difficulty of replacing missing data values is due to the fact that, unlike statistical analysis, DEA is based on a single set of data for each attribute.

The problem of missing data is well recognized in the DEA literature and therefore various approaches for mitigating this issue have been discussed. One such approach is the exclusion of DMUs with missing data from the DEA analysis (Kuosmanen, 2002). This approach has an ill-effect on the efficiency score of the other participating DMUs and may disturb the statistical properties of the estimators. The exclusion of DMUs decreases the production possibility set and increases the efficiency scores of the other units, and may even affect the ranking order of the DMUs being studied. An alternative mitigation approach is the use of dummy values such as zero for replacing the missing output values and a large number for replacing the missing input values. This approach can be