# DATA ENVELOPMENT ANALYSIS WITH SPARSE DATA

by

DEEP KUMAR GULLIPALLI

B.E., Andhra University, India, 2008

A THESIS

submitted in partial fulfillment of the requirements for the degree

MASTER OF SCIENCE

Department of Industrial and Manufacturing Systems Engineering
College of Engineering

KANSAS STATE UNIVERSITY
Manhattan, Kansas

2011

Approved by:

Major Professor
Dr. David Ben-Arieh

# Copyright

# Abstract

Quest for continuous improvement among the organizations and issue of missing data for data analysis are never ending. This thesis brings these two topics under one roof, i.e., to evaluate the productivity of organizations with sparse data. This study focuses on Data Envelopment Analysis (DEA) to determine the efficiency of 41 member clinics of Kansas Association of Medically Underserved (KAMU) with missing data. The primary focus of this thesis is to develop new reliable methods to determine the missing values and to execute DEA.

DEA is a linear programming methodology to evaluate relative technical efficiency of homogenous Decision Making Units, using multiple inputs and outputs. Effectiveness of DEA depends on the quality and quantity of data being used. DEA outcomes are susceptible to missing data, thus, creating a need to supplement sparse data in a reliable manner. Determining missing values more precisely improves the robustness of DEA methodology.

Three methods to determine the missing values are proposed in this thesis based on three different platforms. First method named as Average Ratio Method (ARM) uses average value, of all the ratios between two variables. Second method is based on a modified Fuzzy C-Means Clustering algorithm, which can handle missing data. The issues associated with this clustering algorithm are resolved to improve its effectiveness. Third method is based on interval approach. Missing values are replaced by interval ranges estimated by experts. Crisp efficiency scores are identified in similar lines to how DEA determines efficiency scores using the best set of weights.

There exists no unique way to evaluate the effectiveness of these methods. Effectiveness of these methods is tested by choosing a complete dataset and assuming varying levels of data as missing. Best set of recovered missing values, based on the above methods, serves as a source to execute DEA. Results show that the DEA efficiency scores generated with recovered values are close within close proximity to the actual efficiency scores that would be generated with the complete data.

As a summary, this thesis provides an effective and practical approach for replacing missing values needed for DEA.

# Table of Contents

# List of Figures

# List of Tables

# Acknowledgements

The completion of this thesis will be a milestone in my life. I have to accept that the road to reach here requires enormous amount of love and strength from family and friends to overcome the obstacles. I take this as stage to thank all the people who made my journey more eventful for the last three years at Kansas State.

My deepest gratitude goes to my advisor Dr. David Ben-Arieh for bringing out the best of me. He showed the path to identify bigger picture of this research, and always available to discuss and share ideas. I really appreciate all his support, patience, and perseverance he provided to finish this thesis on time.

I would like to thank Dr. John Wu for teaching Advanced Linear Programming class on special request out of his busy schedule. This helped me greatly to understand the in-depth nature of linear programming and critical theorems. I would also like to thank Dr. Paul Nelson for serving on my committee and for his valuable suggestions to complete this thesis.

I will be forever indebted to my parents for their unconditional love, endless support, and persistent confidence in me, which made my shoulders feel light. They are my first teachers who taught me to strive for excellence in everything that I pursue.

Words will fail, if I have to express thankfulness to my friends. Space will become limited, if I have to mention all of them. This thesis is incomplete without them.

# Dedication

With Love to my Parents

Venkateswara Rao Gullipalli & Vara Lakshmi Gullipalli

Sister - Chandrika Gullipalli

# Chapter 1 - INTRODUCTION

This chapter introduces the motivation, research objectives and contributions to develop new effective methods to estimate missing values in a dataset and then to carry out Data Envelopment Analysis (DEA). The focus of this thesis is to determine the productivity of 41 member clinics of Kansas Association of Medically Underserved (KAMU), using the available sparse data. Traditionally complete data should be available to carry out the Data Envelopment Analysis. Most of the real world cases will not be able to meet this requirement. This hinders the robustness of DEA methodology. The ability to estimate missing values precisely improves robustness of DEA methodology and also the accuracy of the results, since it is susceptible to missing data. This chapter also presents essential assumptions, overview of research results, and outline of this thesis. This research effort intersects with other research domains such as the concept of correlation, fuzzy clustering, and interval range to estimate missing values. Based on these domains three different approaches are proposed to estimate missing values.

DEA is a linear programming methodology and also a non-parametric approach to evaluate relative technical efficiency of homogenous Decision Making Units (DMUs), using common multiple inputs and outputs. DMU can be defined as an organization or business process which consume resources and produce goods or services. DMU can be for-profit or a non-profit organization. This means a hospital, for-profit or non-profit, can be considered as a DMU since it consumes the resources such as medical staff (doctors and nurses) to treat patients, and generates revenue.

DEA methodology determines the best set of weights for multiple inputs and outputs considered for each DMU, using linear programming methodology, to bestow the target DMU with best efficiency score. The efficiency score is calculated as the ratio of weighted sum of outputs to weighted sum of inputs. DEA is very suitable to be applied in healthcare since healthcare providers can be easily identified as DMUs. DEA effectively handles the multiple input and output parameters involved in a healthcare environment. DEA is a non-parametric methodology which does not require prior relationship or functional form between inputs and outputs. Literature review, Chapter 2, provides greater detailed information about DEA, advantages and disadvantages, and its application in healthcare.

This chapter is structured as follows. Section 1.1 represents the motivation for this research work which includes the alarming rise in healthcare expenditure and also the influence of missing data on outcomes of research endeavors. Section 1.2 portrays the research objectives and contributions, and brief introduction to the three different methods. The assumptions developed to ensure the effectiveness of results are presented in Section 1.3. The overview of research results and adopted procedures are revealed in Section 1.4. Finally outline of this thesis work is exposed in Section 1.5.

## 1.1 Motivation

This section presents two important reasons to carry out this research. The first one is the alarming rise of healthcare costs. In order to reduce these costs we need to identify productivity levels, which provides the opportunity for continuous improvement. DEA methodology is widely recognized as an effective methodology to identify relative efficiency scores since its inception. The other reason is that DEA is vulnerable to missing values. In order to improve the robustness of this methodology we need to estimate the missing values more precisely.

### *1.1.1 Rise in Healthcare Expenditure*

As per the encyclopedia of public health (Kirch, 2008), healthcare is defined as the prevention, treatment, and management of illness and the protection of mental and physical well-being through the services provided by the medical nursing, and allied health professions. Healthcare industry is considered as one of the largest industries in the world. It is also the fastest growing industry consuming almost 10% of the Gross Domestic Product (GDP) in most developed nations. The amount of public and private money spent on healthcare services in a country at a given time indicates the country's health expenditure. As per World Health Organization (WHO) the total expenditures on health by United States of America (USA) as a percentage of its GDP for the years 2003 to 2009 are 15.2%, 15.4%, 15.2%, 15.3%, 15.7%, 15.2% and 16.2% respectively. The average health expenditure for other regions of the world was around 10% of their GDP by end of 2008. One can clearly identify health expenditure in USA as an outlier when compared to other regions of the world.

Center for Medicare & Medicaid Services (CMS) is a United States federal agency which administers Medicare and Medicaid programs. As per CMS the National Health Expenditures (NHE) in billions of dollars for the years 2003 to 2009 are $1,772.2, $1,894.7, $2,021.0,

$2,152.1, $2,283.5, $2,391.4, and $2,486.3 respectively. If this situation continues then USA's health care expenditure is expected to increase much faster than the overall economy. The projected forecast summary of CMS signifies that the National Health Expenditure in United States is expected to reach $4.6 trillion, and which accounts for 19.8 percent of the GDP by the year 2020.

The United Stated National Health Expenditures (NHE) in billions ($) and its share of GDP, as per the statistics of CMS are shown in Figures 1.1 and 1.2 respectively.

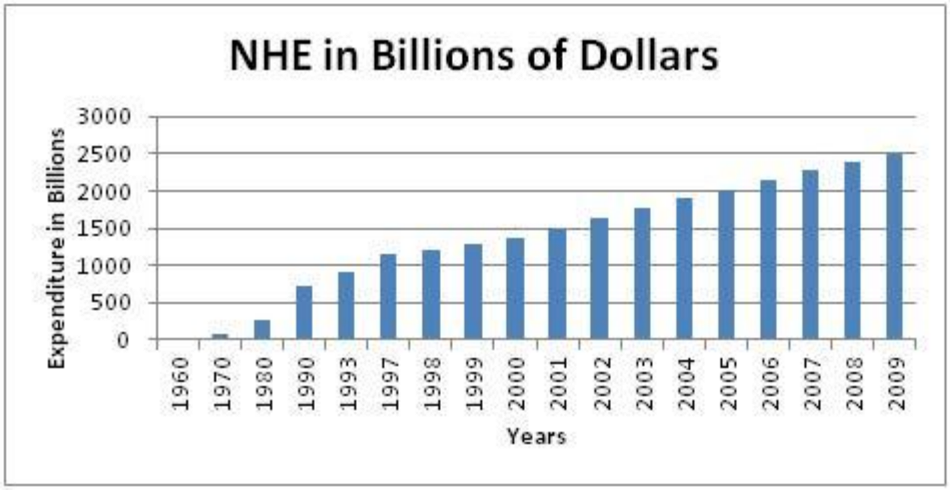

**Figure 1-1: National Health Expenditure (NHE) in Billions ($) from 1960 to 2009**

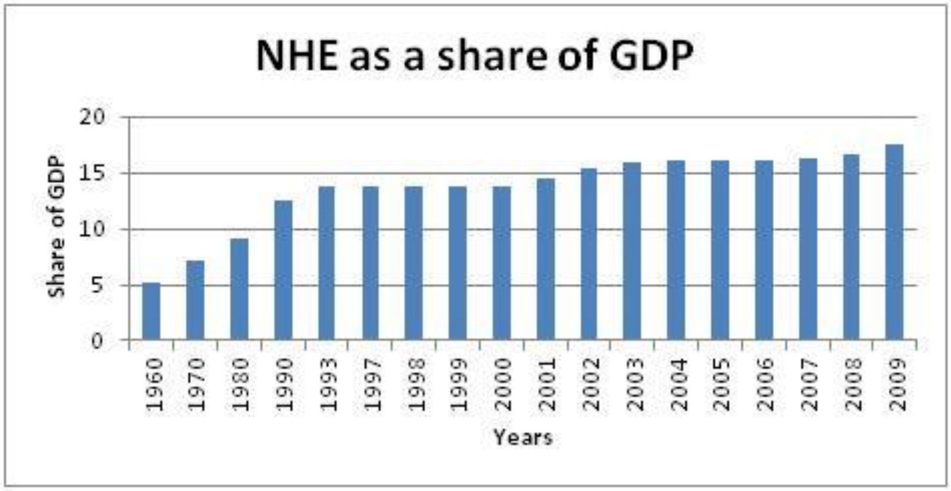

**Figure 1-2: National Health Expenditure (NHE) as a share of GDP from 1960 to 2009**

An alarming increased projection in health care expenditure has driven the administration in search of effective methods to reduce costs associated with healthcare. Organizations such as Agency for Healthcare Research and Quality (AHRQ), Center for Disease Control and Prevention (CDC), Institute for Healthcare Improvement (IHI), and National Association for Healthcare Quality (NAHQ) have taken initiatives to determine effective ways to deliver high quality healthcare at competitive cost through research findings. The following is a minute list of summaries of valuable research findings to reduce healthcare expenditure.

**Implementation of Lean:** Lean system developed by Toyota to remove waste, activities or services which don't add value to the process, is an effective and adaptable approach to reduce waste and inefficiency in healthcare process also through lean redesign. The common forms of waste that usually occur in a manufacturing environment can also be observed in a healthcare environment. The most frequently used activities of lean are Value Stream Mapping, 5S, and Kaizen events to envision and eliminate wastes. A framework of factors that usually affect the implementation of lean process is developed. Reports suggest that lean was implemented in more than 50 locations at every geographic location in United States (AHRQ).

**Healthcare Information Technology**: "Evidence on the Costs and Benefits of Health Information Technology", published by the United States Congress in 2008 discusses benefits of Information Technology in Healthcare. Electronic Medical Record (EMR) is the most common IT package used by healthcare providers to deliver effective healthcare and to reduce the physical efforts of physicians, duplication, and medical transcription errors. Studies have estimated that $80 billion could be saved in net annual due to such technology, but the fact is that only 12% of physicians and 11% of hospitals have adopted it as of 2006. Adoption of such IT health systems can create both internal and external savings. Evidences also suggest that such IT health systems can improve productivity of nurses and physicians, and can reduce average length of patient stay in the hospital by 5%, through speeding up certain hospital functions.

**Benchmarking:** Benchmarking is considered as the process of comparing the performance metrics of a particular process/product/service/organization with the best standards in that particular area. This helps in identifying the targets/projections/practices required to be the best in that area. Benchmarking is viewed as a continuous improvement tool providing opportunity for organizations to be best in their class. The functional application of benchmarking is not limited to any particular industry. The early use of benchmarking in National Health Services (NHS) identified by Pantall (2001) can be dated back to 1960. He also provided particulars on how organizations such as IBM found benchmarking as an effective tool for sharing best practices with other organizations and how it evolved as an important tool within continuous improvement methodologies. Healthcare industry also considered benchmarking as a simultaneous opportunity to reduce expenditure and improve the quality of healthcare. For further discussion on different type of benchmarking techniques refer to Benson (1994).

This thesis considers benchmarking as an effective tool to measure the performance of 41 member clinics of Kansas Association for the Medically Underserved (KAMU), and to share the best practices. This thesis considers Data Envelopment Analysis (DEA), a relative technical efficiency measurement technique, to identify the benchmarks. Based on the results of DEA the clinics can be classified into three different groups. The first group consists of a group of super clinics that are very productive and efficient in using their resources. A second group consists of clinics that are quite effective and the third group contains clinics that consume more resources than other clinics and produce less outputs.

KAMU is a non-profit organization founded in 1989 and was recognized as the Primary Care Association (PCA) of Kansas in 1991. As a PCA, it is charged with providing training, technical assistance and advocacy on behalf of federally funded Community Health Centers in Kansas. Its mission is to support and strengthen its member organizations through advocacy, education and communication (KAMU). More information about KAMU and preparing the data for Data Envelopment Analysis can be found in Chapter 3.

### *1.1.2 Missing Values in Healthcare*

Missing values occur when any required data value for any observation or variable is either not recorded or misplaced during transfer of data. The other type of absentness in data occurs when the individuals are reluctant to provide data. Either ways the missing values can greatly influence the results of research efforts. In most cases it is common to find data with missing values. The missing values occur due to technical errors, i.e., breakdown of machine, human errors, i.e., fail to record or entry of data, natural calamities, i.e., bad weather.

The nature of missing values can be classified into three different groups Missing Completely at Random (MCAR), Missing at Random (MAR), and Missing Not at Random (MNAR), Little and Rubin (2002). Missing Completely At Random (MCAR) implies that data found to be missing does not follow any particular reference to either the data values present or missing. Items found to be missing has equal probability to be missing. Missing At Random (MAR) implies that data found to be missing can be random, but the nature of missing values can be attributed to a particular reason. If the missing data cannot be classified into above two groups, MCAR and MAR, then it belongs to Missing Not At Random. The values found to be missing are biased towards particular reason. For more detailed information on missing data and for statistical analysis with missing data refer to Little and Rubin (2002).

Missing values are part of almost every domain and research work; healthcare and data envelopment analysis are no exceptions to them. Norris et al., (2000) identifies that missing values in clinical registry is a common phenomenon and can affect any research outcomes and analyses. Faris et al., (2002) compared 3 multiple imputation methods to enhance a clinical database with missing values. Graham (2009) reviews strategies based on strong statistical traditions, clearing the myths and misconceptions, to make missing data analysis methods useful in the real world. Most of the methods in literature are based on statistics to estimate missing values.

Data Envelopment Analysis is easily susceptible to missing data since it depends on single dataset of chosen inputs and outputs, unlike statistical methods. The most common methods to deal with missing values such as list-wise deletion and pair-wise deletion cannot be applied to DEA, since it reduces the total number of DMUs to perform benchmark analysis and reduces the sample size. Efficiency scores of the DMUs in a group can be greatly influenced by reducing the number of DMUs; since DEA efficiency scores are relative to the DMUs in that group. In order to surmount the issues of adequate quantity of data for DEA the missing data needs to be estimated to the most possible accurate level.

In this thesis we focus on development of new reliable methods to deal with missing values in DEA and thereby to improve robustness of this methodology. These methods are applicable to other analyses with missing values but they are specifically developed with a view point for DEA. This thesis work provides an effective and practical approach for replacing missing values needed for a DEA analysis. The disadvantages of DEA with missing data can be compensated by taking proper care about the quality and quantity of data.

## 1.2 Research Objectives & Contributions

### *1.2.1 Research Objectives*

This main focus of this thesis is development of new reliable methods to handle Data Envelopment Analysis with missing data. The associated research objectives of this thesis are as follows:

- Productivity measurement of 41 member clinics of Kansas Association for the Medically Underserved (KAMU) with sparse data. Determination of benchmarks among them for sharing best practices with other organizations.
- Identify what areas need to be improved for each clinic, and provide quantitative guidelines to achieve the best standards.
- Study of existing methods to handle the issue of missing data in DEA.
- Development of new reliable methods to estimate the missing data and understanding the effect of these recovered values on DEA scores.
- Illustration of the developed methods using example datasets.
- Application of these developed methods on KAMU clinics.
- Evaluating the effectiveness of these methods by comparing the results with those of existing methods.
- Identification of the limitations for these methods so that they provide results with greater precision.

### *1.2.2 Research Contributions*

The research efforts for this thesis intersect with other research domains such as the concept of correlation, fuzzy clustering, and interval approach to estimate missing values. Based on these domains three different methods are proposed to estimate the missing values. The contributions made for these methods are as follows.

First method named as Average Ratio Method (ARM) uses the average value of all the ratios between two variables. The precision to estimate the missing values depends on the amount of correlation between two variables; greater the correlation greater the accuracy of results. The selection procedure for such variables and step by step procedure of this method will be addressed in Chapter 4. The advantages of this method are it is less computational, and produces better set of results when compared to other basic methods. The limitations of this method are it requires additional data with good correlation. Effectiveness of this method is tested by comparing it with other basic methods form the literature, using the example datasets provided in the literature works.

Second method is based on the concept of modified fuzzy c-means clustering algorithm which can handle missing values, an existing algorithm. We identified that this particular algorithm developed by Hathaway and Bezdek (2001) is susceptible to two major issues. One, the missing values in the data needs to be substituted by some initial values prior to beginning of the algorithm. This particular algorithm is sensitive to such values chosen initially. Two, it is also susceptible to the number of clusters to be chosen. This research effort addresses these two major issues to improve the effectiveness of this algorithm. Greater details of the modified fuzzy c-means clustering algorithm and description of research endeavors is presented in Chapter 5.

The other major issue associated with the modified fuzzy c-means algorithm identified by Himmelspach and Conrad (2010) is cluster dispersion. Cluster dispersion reduces the likelihood of remote data objects being biased by cluster size. The three newly developed approaches which try to achieve similarity among the cluster sizes reducing the opportunity for formation of few large cluster groups will be discussed in Chapter 6.

Third method is based on interval approach to handle the issue of missing values and to perform Data Envelopment Analysis. Missing values are replaced by interval ranges estimated by experts or based on statistical techniques, rest remains crisp. In most interval based DEA methods, the efficiency scores are expressed in terms of fuzzy environment but in this case the scores are expressed as crisp values. Crisp efficiency scores are identified in similar lines to how DEA determines efficiency scores of DMUs using the best set of weights. Intervals are split into crisp values based on linear interpolations, using common value of alpha. Best value of alpha, for interval ranges, will be one which endows most of the DMUs with best efficiency scores, further insight into this method will be presented in Chapter 7.

These approaches are demonstrated using the real and complete dataset of 22 KAMU clinics, assuming varying levels of data as missing. Insight to the identification procedure of complete dataset from the sparse dataset will be provided in Chapter 3.

## 1.3 Assumptions

Apart from the basic assumptions of DEA which will be presented in the literature review, Chapter 2. This section presents some important hypothesis. The following are considered to be the basic assumptions to determine the productivity of KAMU clinics:

- All the clinics are assumed to be within patients reach, in order to nullify the influence of geographic nature on performance measurement
- All the clinics are assumed to be functioning for the same number of days and effectiveness of the technical and administrative staff is considered to be equivalent over all clinics
- All the clinics are assumed to have similar kinds of services

We have made these assumptions since they can be critical parameters for performance measurement and can be influential in predicting DEA results. Even though DEA has a potential ability to address these kinds of issues, they are not primary concerns to this thesis. Methods to effectively handle missing data to execute DEA are our primary objectives. The literature review, Chapter 2, guarantees on how these issues can be addressed.

The productivity measurement study of 41 member clinics of KAMU does not try to justify or explain the differences between the clinics which can very well be justifiable.

## 1.4 Research Results

This section presents an overview of the adopted procedures and research results which conveys the significance of this research work and effectiveness of the newly developed methods.

There is no particular methodology to determine the effectiveness of methods which estimate the missing values. In real cases the missing value is never known, in order to estimate the effectiveness of these methods we have to assume known value as missing. The usefulness of the methods developed in this thesis are tested based on the real and complete dataset of 22 clinics, chosen from the KAMU sparse data. Assuming varying levels of the data as missing for different nature of missing values, the effectiveness of the methods presented in this thesis are judged.

The percentages of missing values in the data are varied from 10% to 40%. The three different nature of missing values MCAR, MAR, and MNAR are considered to test the effectiveness. The effectiveness of these methods is tested based on the ability to recover the missing values within closeness to the assumed missing values (known and real). The difference between the real (assumed as missing) and recovered values (estimated using the methods) is determined using the Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD). The best set of recovered values serve as a source to carry out the Data Envelopment Analysis. In most cases the methods are able to estimate within close proximity. Few methods are also tested based on the data obtained from the literature.

## 1.5 Outline of Thesis

Chapter 2 reviews all the important concepts that are germane to Data Envelopment Analysis. It describes the background, evolution of DEA, basic models of DEA. Assessment of strengths and limitations of DEA is carried out and then this chapter digs into the literature works on evaluation of efficiency measures in healthcare using DEA.

Chapter 3 introduces the mission, objectives, and programs of Kansas Association for the Medically Underserved. It also provides an overview of the data provided by KAMU. This chapter primarily focuses on issues associated and measures need to be taken for preparing the data for Data Envelopment Analysis. The literature review at the end of the chapter addresses the methods to handle the missing data during Data Envelopment Analysis.

Chapter 4 introduces the Average Ratio Method (ARM) methodology to determine the missing values. Greater details about this methodology, step by step procedure, its advantages, and limitations will be discusses in this chapter. The proposed methodology is used to evaluate the efficiency of 41 KAMU clinics with sparse data

Chapter 5 presents a methodology based on fuzzy clustering concepts to execute Data Envelopment Analysis with sparse data. It provides an introduction to data clustering, then to fuzzy clustering concepts. The issues associated with the existing algorithm are eliminated to improve its effectiveness. These suggested approaches are demonstrated on a real and complete dataset of 22 KAMU clinics, assuming varying levels of missing data.

Chapter 6 presents a methodology to estimate missing values based on a modified fuzzy c-means clustering algorithm which takes cluster dispersion into account. The reasons behind the failure (does not converge) of existing cluster dispersion method are illustrated. New cluster dispersion approaches are proposed in this chapter. The newly developed clustering approaches are demonstrated on a real and complete dataset of 22 KAMU clinics, assuming varying levels of missing data.

Chapter 7 presents an interval approach based methodology to handle the issue of missing values and to perform Data Envelopment Analysis. Missing values are replaced by interval ranges estimated by experts, rest remains crisp. The primary focus of this methodology is to determine the crisp efficiency scores out of interval ranges. This new approach is demonstrated on a real and complete dataset of 22 KAMU clinics, assuming varying levels of data as missing.

Chapter 8 is the final chapter which contains the summary of the research efforts, its outcomes, and scope for future research.

# Chapter 2 - LITERATURE REVIEW

This chapter reviews all the important concepts that are germane to Data Envelopment Analysis (DEA). This includes basic definitions to understand the concept of productivity, different measures to evaluate productivity, and their advantages and disadvantages. Then this chapter describes the background and evolution of DEA, and graphically illustrates DEA methodology using an example. This chapter also introduces all the basic models of DEA, their significance, and variation between different models. There are several other models developed in DEA since its inception, other than the basic models, which are generally applied based on the requirements. Assessment of strengths and limitations of DEA is carried out and then this chapter digs into the literature works on evaluation of efficiency measures in healthcare using DEA. Summary of selected literature works are presented in the final section of this chapter.

This chapter is structured as follows. Section 2.1 presents the basic concepts such as productivity, relative technical efficiency, production function and frontier, and economic returns to scale. Section 2.2 introduces the different efficiency measurement techniques by digging into the literature and also provides advantages and disadvantages for each method. Section 2.3 graphically illustrates the methodology of DEA, while section 2.4 provides the complete details since inception of DEA. It also introduces the two basic methods of DEA, their formulation, primal and dual approach, input and output orientations. Section 2.5 introduces other important models of DEA and their specific usage. Strengths and limitations of DEA are presented in section 2.6. Finally literature review on application of DEA in healthcare and the need for efficiency measurement techniques is presented in Section 2.7

## 2.1 Basic Definitions

### *2.1.1 Productivity*

Prokopenko (1987), defined productivity as the efficient use of resources consumed for production of various goods and services. It develops the relationship between outputs produced to inputs consumed for the production. High value of productivity implies high capacity to achieve greater outputs with same quantity of inputs or to achieve the same volume of outputs with lesser quantity of inputs.

$$Productivity = \frac{Output}{Input}$$

Productivity measures serve as a comparative tool helping employees at various levels of organizations to evaluate the effectiveness of process/products. Such measures can be either partial measure or total factor measures, which will be discussed in the coming sections. Farrell (1957) extended the concept of productivity to a more general concept called efficiency, which involves technical efficiency and allocative efficiency.

### *2.1.2 Technical Efficiency*

It is one of the two measures proposed by Farrell (1957) to measure the efficiency of a firm. It can be defined as the situation under which the firm cannot produce more amount of output for given available input resources and also the firm cannot produce same amount of output with less amount of available input resources. It can determine the amount of waste that can be eliminated without worsening any input or output. Mathematically technical efficiency of an organization is attained when the two following conditions are satisfied.

1. $\theta^* = 1,$ When the optimal efficiency of an organization is 100%
2. $s^{-*} = s^{+*} = 0,$ When the input and output slacks of an organization are zero

### *2.1.3 Allocative Efficiency*

Farrell’s second measure of efficiency is also known as Price Efficiency. Allocative efficiency is considered when information related to prices, cost minimization, and profit maximization is available. Allocative efficiency can be attained when the organization is technically efficient and is able to achieve it at a minimum total cost of production. It can be defined as a situation when price of goods or services are closer to the marginal value of the resources used for production.

*Overall Efficiency = Allocative Efficiency * Technical Efficiency*

Farrell’s two efficiency measures, Technical and Allocative Efficiency, can together provide the overall economic efficiency.

### *2.1.4 Production Function*

Farrell (1957) introduced the concept of efficient production function to provide a satisfactory measure of efficiency. Efficient production function is the maximum amount of output that can be obtained from any combination of inputs. The two possible options to construct the efficient production function are either based on a theoretical function or an empirical function. It is very difficult to develop a theoretical function for a typical complex firm like a manufacturing industry since some problems might be overlooked. As the complexity of the function increases, accuracy of the results decreases. On the other hand empirical function estimates the efficient production function based on the observation of inputs and outputs of a number of firms. Farrell justifies the use of empirical function by saying that “it is far better to compare performance with the best actually achieved than with some unattainable ideal.”

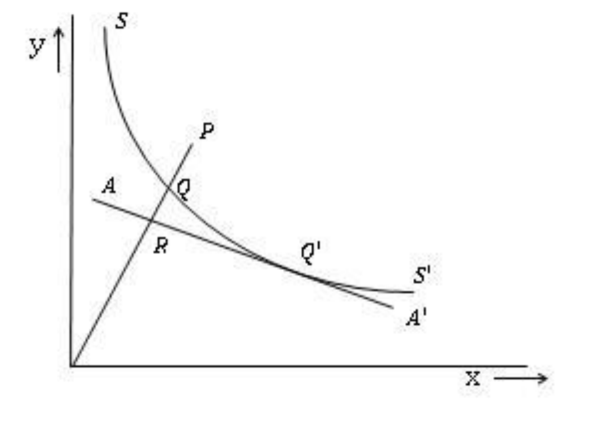

**Figure 2-1: Isoquant Diagram**

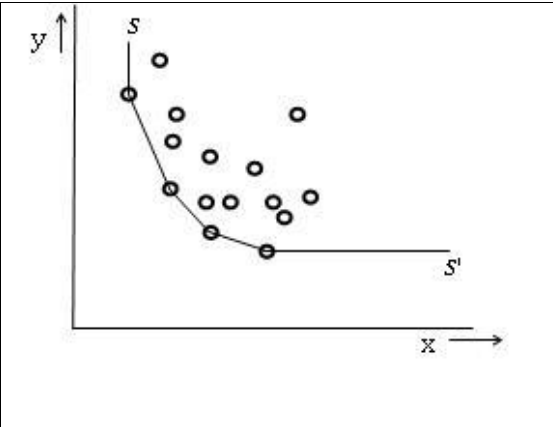

**Figure 2-2: Production Frontier**

Farrell explains the concept of efficient production function using a simple example. Consider a firm using two inputs to produce a single output under the assumption of Constant Return to Scale (CRS). CRS means that outputs vary by the same proportion as inputs, further details will be discussed in the coming section. The example is represented by isoquant as shown in Figure 2.1.

Isoquant curve represent all possible combinations for either inputs or outputs that define the production function at a constant level of outputs or inputs respectively. The Isoquant curve shown in Figure 2.1 represents the possible combinations of inputs for constant value of outputs. Such an Isoquant curve is known as input orientation. Isoquant curve $SS'$ represents the set of points which can produce the same quantity of output for different combinations of inputs. Each point on the Isoquant curve represents a production unit. Figure 2.2 represents each firm as a point and different firms are represented by the scatter plot.

The two important assumption made by Farrell to make this possible are

1. The isoquant is convex to the origin. It means that if two points are attainable in practice then so is their weighted average.
2. The slope of the isoquant is nowhere positive. It is to ensure that an increased application of both inputs will not result in reduced output.

Isoquant curve $SS'$ represent various combinations of inputs consumed by an efficient firm to produce a unit output. Point $Q$ which lie on the Isoquant curve $SS'$ represents an efficient firm. Point $P$ represents an inefficient firm using the same proportion of inputs as firm $Q$. Both firms $P$ and $Q$ produce same quantity of output, but firm $Q$ uses only fraction $OQ/OP$ for each factor consumed by $P$. It can also be explained as firm $Q$ could produce $OP/OQ$ times more amount of output for same quantity of inputs. The ratio $OQ/OP$ is defined as the **technical efficiency** of firm $P$.

Observation of points $Q$ and $Q'$ reveals that these points are technical efficient because they lie on the isoquant. The firm $Q'$ is the optimal method of production but not $Q$. The cost of production at $Q'$ will only be a fraction $OR/OQ$ of those at $Q$. The ratio $OR/OQ$ is defined as the **price efficiency** of the firm $Q$. This ratio can also be considered as the price efficiency of firm $P$. As the firm $P$ tries to reach its proportion of inputs as $Q'$, in order to reach technical efficiency, its costs need to be reduced by the factor $OR/OQ$.

The product of technical efficiency and price efficiency produces the **overall efficiency** of the firm. The ratio $OR/OP$ is defined as the overall efficiency of the firm $P$. Farrell (1957) outlined that technical efficiency measure success by producing maximum amount of output for the same amount of input. Price efficiency measure success by choosing optimal set of inputs to identify optimal method of production.

### *2.1.5 Production Frontier*

Production frontier is a more general concept than production function. DEA is the outcome of linking Farrell’s technical efficiency concept with production frontier. Production frontier represents the list of all efficient firms, which can attain maximum output level for a given input level. A firm is considered to be technically efficient if it lies on the production frontier. Firms lying outside the production frontier are considered to be inefficient. The points lying on the isoquant $SS'$ represent efficient firms and points lying at a distance away from frontier represent inefficient firms as shown in Figure 2.2.

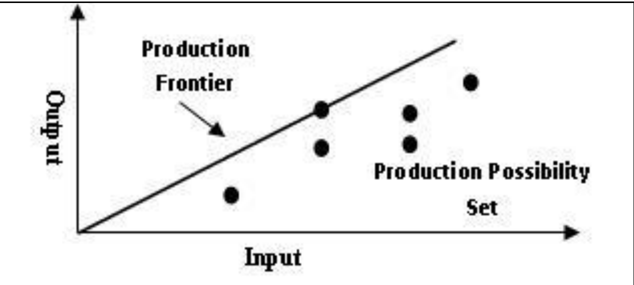

**Figure 2-3: CCR Production Frontier**

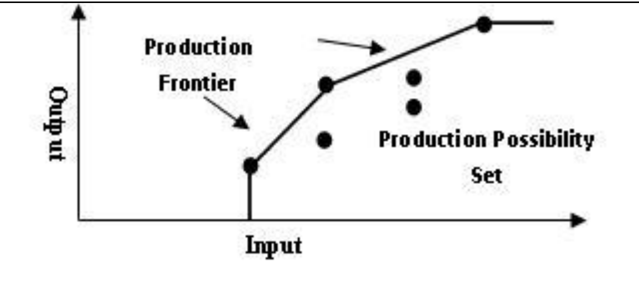

**Figure 2-4: BCC Production Frontier**

Production frontiers developed in DEA are based on the non-stochastic methods. There exist methods to determine the productivity using the stochastic methods, which will be addressed in later sections. Production frontiers developed in DEA are not ideal frontiers. They are developed based on sample data provided by the firms. Characteristics of production frontiers can vary based on returns to scale (RTS), which will be discussed in the next section. Charnes, Cooper, and Rhodes (CCR) model assumes Constant Returns to Scale (CRS), so the production frontier will be linear. Banker, Charnes, and Cooper (BCC) model assumes Variable Returns to Scale (VRS), so its production frontier is formed by the convex hull. Inefficiency of the firm in both cases is determined by projecting the inefficient firm onto the frontier. More details about the CCR and BCC models will be presented in section 2.4. The production frontier for CCR and BCC models are shown in Figure 2.3 and 2.4 respectively.

### *2.1.6 Returns to Scale*

Returns to scale describes the change in output scale of production in long run for change in input levels. The different returns to scale are:

#### *2.1.6.1 Constant Returns to Scale (CRS)*

The first DEA model by Charnes, Cooper, and Rhodes (CCR) is based on the concept of constant return to scale. For the proportionate change in all inputs, if all outputs vary by the same proportion then the production function exhibits constant returns to scale, Coelli (2005). For example consider a firm producing single output using single input, say number of employees. The production is expected to double if the number of employees is doubled. Mathematically if all the inputs are scaled by an amount k > 1, then

$$f(kx) = kf(x)$$

#### *2.1.6.2 Variable Returns to Scale (VRS)*

The DEA model by Banker, Charnes, and Cooper (BCC) is based on the concept of variable returns to scale. If for the proportionate changes in all inputs the output results vary by a different proportion, then the production function exhibits Variable Returns to Scale, Coelli (2005). Variable returns to scale can be further classified as Increasing Returns to Scale (IRS) and Decreasing Returns to Scale (DRS).

#### *2.1.6.2a Increasing Returns to Scale (IRS)*

If the outputs vary by a proportion greater than the proportion of inputs then the production function exhibits IRS. Mathematically if all the inputs are scaled by an amount k > 1, then

$$f(kx) > kf(x)$$

#### *2.1.6.2b Decreasing Returns to Scale (DRS)*

If the outputs vary by a proportion lesser than the proportion of inputs then the production function exhibits DRS. Mathematically if all the inputs are scaled by an amount k > 1, then

$$f(kx) < kf(x)$$

## 2.2 Efficiency Measurement Techniques

The efficiency measurement techniques can be generally classified into two groups Partial Productivity measures and Total Factor Productivity measures (TFP). Partial productivity measures develop a ratio between a single input and output. Average labor productivity is the most commonly used partial productivity measure, which evaluates the output per worker employed. Other similar partial productivity measures are fuel productivity in power stations and land productivity in agriculture, Coelli (2005). The limitations of partial productivity measures are they can provide false information, since they fail to account the influence of other resources on productivity. Example, gain in productivity either due to machinery or management changes might be attributed to labor hours, Cooper et al., (2007). Total factor productivity measures take multiple inputs and outputs into account to determine productivity of a firm. The difficulties associated with total factor productivity measures are choice of inputs and outputs and assignment of weights. Initially there exists contrast between the usage of fixed weights and variable weights, chosen based on a best set, for each entity to be evaluated. DEA uses the concept of best set of weights to determine the efficiency scores. The following are different efficiency measurement techniques.

### *2.2.1 Ratio Analysis*

Ratio analysis is one of the most commonly used early techniques by analysts to evaluate performance of banks. It is a powerful tool for financial analysis. Ratios identify the relationship between two variables and helps in simplifying the information of financial statements. Any number of ratios can be designed to compare the performances between banks and its branches over a period of time, Siddiqui (2005).

Ratio analysis seems to be simplistic in providing the information but the complexity of it increases as the number of ratios keeps increasing. The concept of unlimited number of ratios is often contradicting and confusing. This approach limits the productivity measure to single input and output; it cannot be extended to multiple inputs and outputs. It does not acquire the competence for identifying inefficient firms and predicting the projections required for their performance improvement, Paradi et al., (2004).

### *2.2.2 Indices of Efficiency*

Productivity measures based on single criterion were not satisfactory and efforts to identify the measures which consider multiple factors are persistent. Indices of efficiency is one such attempt to measure the efficiency by adding up different factors of a firm. The indices of efficiency sought to represent the dimensionless input quantities by weighted averages. The weighted average is equivalent to the valuation of inputs and price proportional to the weights.

This simplifies it to a cost comparison, if all the firms in the analysis choose the same set of prices. The difficulty arises in choosing a suitable set of weights; otherwise the choice of a set of prices should be subjective, Farrell (1957).

### *2.2.3 Regression Analysis*

Regression analysis is a statistical approach with capability to handle multiple inputs and outputs to estimate the relationship between variables. It identifies the average behavior among the variables and also identifies the inefficient units based on the distance from the central tendency of the units.

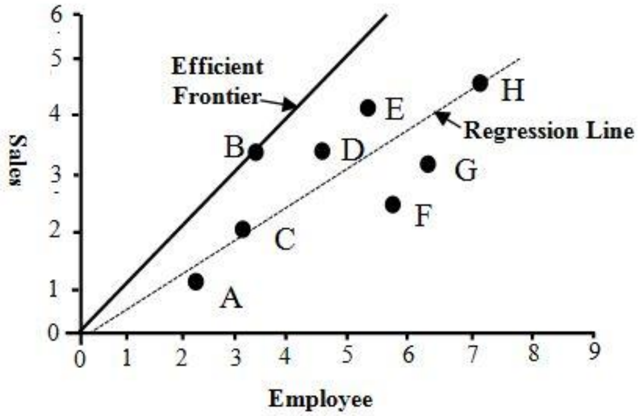

**Figure 2-5: Regression vs. Frontier**

However, it is unable to identify the potential efficient units and the relationship between them. It can seek the units suggesting the need for improvement. Also it cannot determine the required inefficiency area of the firm and the projections required to be efficient. The frontier analysis which will be discussed in the next section possesses these advantages compared to regression analysis; the difference between them is shown in Figure 2.5.

Regression analysis is a complex process to assess the performance using multiple inputs and outputs. The advantage of regression analysis is it can account for random noises in the input and output levels of data, Thanassoulis (1993).

### *2.2.4 Frontier Analysis*

Frontier analysis is a modern day efficiency measurement technique which measures the efficiency of firms based on the distance from the frontier formed by efficient firms. The frontier is developed empirically based on the dataset provided by firms, and we already discussed the issues associated with development of theoretical frontier. Frontier analysis not only estimates the efficiency scores of the production units but also the inefficiency associated with them. It provides projection scores, for multiple inputs and outputs considered, to improve the efficiency score of inefficient units. Frontier analysis provides the flexibility to determine efficiency score of firms under the assumption of alternative returns to scale such as constant, increasing, and decreasing.

Frontier efficiency measurement techniques can be primarily classified into two groups known as parametric and non-parametric methods. Parametric methods require prior defined relationship between inputs and outputs. Non-parametric methods do not require prior relationship between inputs and outputs. The two most commonly used frontier based methods are Stochastic Frontier Analysis (SFA), and Data Envelopment Analysis (DEA).

#### *2.2.4.1 Stochastic Frontiers Analysis (SFA)*

SFA is a parametric methodology which requires functional form to estimate the frontier. It distinguishes between inefficiency and random error by assuming that they have different distributions. Random error is usually modeled using a standard normal distribution with mean zero. Inefficiency is usually modeled using different distributions such as normal, exponential, and gamma. Technical efficiency is calculated using maximum likelihood estimation function, Berger and Humphrey (1997). SFA has not only been employed to study efficiency of hospitals, but also used to study nursing homes, primary care delivery and pharmacies, Hollingsworth (2003). SFA approach ranks the firm with lower costs for a given set of input prices (but the same output quantities) as more efficient than other firms, Paradi et al., (2004).

#### *2.2.4.2 Data Envelopment Analysis (DEA)*

DEA is a linear programming methodology and also a non-parametric approach to evaluate relative technical efficiency of homogenous Decision Making Units (DMUs), using common multiple inputs and outputs. DMU can be defined as an organization or business process which consumes common resources and produces goods or services. DMU can be for-profit or a non-profit organization. DEA methodology determines the best set of weights for multiple inputs and outputs of each DMU, using linear programming methodology, to bestow the target DMU with best efficiency score. The efficiency score is calculated as the ratio of weighted sum of outputs to inputs. DEA is a non-parametric methodology which does not require prior relationship or functional form between inputs and outputs.

### *2.2.5 SFA vs. DEA*

SFA and DEA are two important methods based on frontier analysis, each of them having their own advantages and disadvantages when compared to the other. SFA is stochastic and parametric, whereas DEA is deterministic and non-parametric. SFA can account for noise in the data by separating inefficiency from random error where as DEA cannot separate it. Accuracy of the results of SFA depends on the functional form whereas DEA is non-parametric and does not require any functional forms.

SFA runs single overall optimization principle for all the firms to estimate their inefficiencies. DEA runs separate optimization principle for each firm to estimate their inefficiencies. Bryce (2000) suggested that choice between these two methods depended on nature of the problem. SFA is more helpful to understand the future behavior of the entire population while DEA is used mostly to eliminate inefficiency of individual units specifically, Chilingerian and Sherman (2004).

Hollingsworth (2003) reviewed application of parametric and non-parametric approaches in measuring efficiencies of healthcare units. He found that almost 50% of the studies are based on DEA alone; more than 80% of the studies used DEA either alone or in combination with some other methods, and only 12% studies used SFA.

## 2.3 Graphical Illustration

This section illustrates the methodology of DEA graphically using a simple example. The graphical illustration with the help of a single input and single output provides a better view of DEA methodology.

Consider a simple example of 10 clinics, with input as number of nurses and output as number of patients. The analysis shows the relationship between patients and nurses. Table 2.1 shows the recorded data for input nurses and output patients, and relation between them. It also shows the efficiency scores determined using the CCR Input Oriented Model of DEA.

**Table 2-1: Single Input and Output**

| DMU | (I) # of Nurses | (O) # of Patients | Patients/Nurse | Efficiency Scores |
|---|---|---|---|---|
| A | 5 | 40 | 8.00 | 0.421 |
| B | 8 | 30 | 3.75 | 0.197 |
| C | 2 | 38 | 19.00 | 1.000 |
| D | 4 | 49 | 12.25 | 0.645 |
| E | 9 | 45 | 5.00 | 0.263 |
| F | 7 | 38 | 5.43 | 0.286 |
| G | 5 | 45 | 9.00 | 0.474 |
| H | 6 | 26 | 4.33 | 0.228 |
| I | 8 | 36 | 4.50 | 0.237 |
| J | 3 | 38 | 12.67 | 0.667 |

Representing the data recorded in Table 2.1on a graph by plotting number of nurses on horizontal axis and number of patients on vertical axis. The slope corresponds to the relationship between patients to nurses and this can be observed from Figure 2.6.

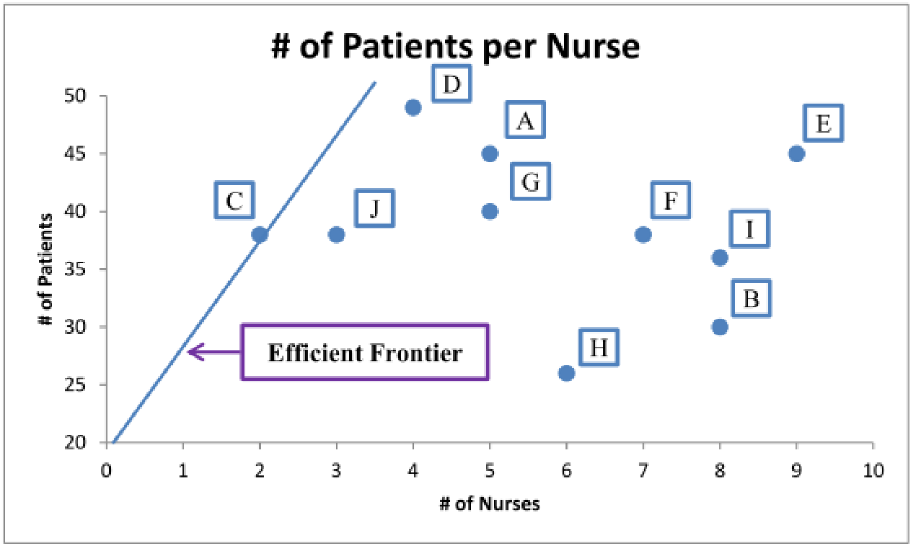

**Figure 2-6: Single Input and Single Output**

The clinic with highest slope forms the efficient frontier and rest of the clinics either lie either on this frontier or below the frontier. Clinics that lie on this frontier are termed as efficient and the one's that lie below the frontier are termed as inefficient. The efficient frontier envelops all the other points in the plane hence it obtained the name Data Envelopment Analysis.

Figure 2.6 shows that clinic *C* is efficient and it forms the frontier inefficient since it possess the highest slope. The other clinics such as *J, D, A, G*, and etc. lies below the frontier are termed as inefficient clinics. Inefficient clinics can be converted into efficient DMU in two ways, either by reducing the amount of inputs consumed or increasing the amount of outputs produced by the clinic. The inefficient clinic *J* can be converted to efficient, if it can reduce the number of nurses from 3 to 2 to treat 38 patients or if it can serve 19 more patients with 3 nurses.

## 2.4 Data Envelopment Analysis

### *2.4.1 Background*

The term Data Envelopment Analysis (DEA) was first brought into use by Charnes, Cooper, and Rhodes in 1978 to evaluate U.S public schools. Their research efforts, of Rhodes under the supervision of Cooper, to evaluate the educational programs for disadvantaged students in a series of large scale studies with support from the federal government is the origin of DEA, Cooper et al., (2004). Rhodes and Cooper started looking into Farrell's, 1957 work on "The Measurement of Productive Efficiency". Charnes was brought in for this topic through his previous research association with Cooper. Charnes, Cooper, and Rhodes extended the germinal ideas of Farrell's research work to develop the basic DEA model. Farrell defined two measures of efficiency in his research work, one is Technical Efficiency, and the other is Allocative Efficiency, which were well defined in the previous sections. DEA can be termed as the extension of Farrell's Technical Efficiency using production function.

### *2.4.2 Terminology*

| | |
|---|---|
| DEA | = Data Envelopment Analysis |
| $DMU$ | = Decision Making Unit, which consume inputs & 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, $\forall\ i = 1, \dots, m$

$u_r$ = Output multiplier variable of ratio model, $\forall\ r = 1, \dots, s$

$x_{ji}$ = Represents input variables of $DMU_j$, $\forall\ i = 1, \dots, m$

$y_{jr}$ = Represents output variables of $DMU_j$, $\forall\ r = 1, \dots, s$

### *2.4.3 Charnes-Cooper-Rhodes (CCR) Model*

CCR model named after Charnes, Cooper, and Rhodes is the first DEA model developed in 1978. CCR model is a fractional programming model which measures the relative technical efficiency of the firms based on multiple inputs and outputs. Efficiency is measured as the ratio of weighted sum of outputs to weighted sum of inputs, Charnes (1978).

$$Efficiency = \frac{Weighted\ Sum\ of\ Outputs}{Weighted\ Sum\ of\ Inputs}$$

Consider a dataset of $n$ DMUs which consume $m$ inputs and produce $s$ outputs. Input and output data for $DMU_j$ are represented as, $x_{ji}$ $(i = 1, \dots, m)$, and $y_{jr}$ $(r = 1, \dots, s)$ respectively, where $(j = 1, \dots, n)$. Efficiency of each DMU is evaluated relative to the constraint set of all $n$ DMUs, and needs $n$ optimizations to evaluate the efficiency scores of all the DMUs. DMU under evaluation is represented by $DMU_o$. The following is the fractional programming model based on the definition of efficiency.

$$Max \quad Z = \frac{\sum_{r=1}^{s} u_r y_{0r}}{\sum_{i=1}^{m} v_i x_{0i}}$$

S.T (1)

$$\frac{\sum_{r=1}^{s} u_r y_{jr}}{\sum_{i=1}^{m} v_i x_{ji}} \leq 1 \quad \forall j = 1, \dots, n$$

$$u_r, v_i \ \geq 0 \ \forall\ r = 1, \dots, s, \quad i = 1, \dots, m$$

Charnes in 1978 converted the Fractional Programming problem model (1) into a linear programming problem model (2). We solve the linear programming problem to obtain values for input weights, $v_i \ (i = 1, \dots, m)$, and output weights, $u_r \ (r = 1, \dots, s)$, as variables which need to satisfy the constraints set and to optimize the objective function. Constraint set restricts the ratio of weighted sum of outputs to inputs to not exceed unity for every DMU. Model (2) is also known as the multiplier approach due to use of input and output multiplier weights.

$$Max \quad Z = \sum_{r=1}^{s} u_r y_{0r}$$

S.T (2)

$$\sum_{i=1}^{m} v_i x_{0i} = 1$$

$$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \ \leq 0 \ \forall \ j = 1, \dots, n$$

$$u_r, v_i \ \geq 0 \ \forall \ r = 1, \dots, s, \quad i = 1, \dots, m$$

$$Min \quad Z = \ \theta$$

S.T (3)

$$\theta x_{oi} - \sum_{j=1}^{n} x_{ji} \lambda_j \ \geq 0, \quad \forall \ i = 1, \dots, m$$

$$\sum_{j=1}^{n} y_{jr} \lambda_j \ \geq y_{or}, \forall \ r = 1, \dots, s$$

$$\lambda_j \ \geq 0 \ \forall \ j = 1, \dots, n$$

Model (3) represents the dual linear programming problem of primal model (2). Primal and Dual are transposition to each other. If primal is a maximization problem then dual will be a minimization problem. Dual is used to determine the amount of inefficiency of DMUs by projecting them onto the efficient frontier. In this case dual aims at minimization of inputs. Model (3) is also known as Envelopment approach, due to formation of envelop to evaluate the inefficiency of DMUs.

Generally dual is referred to as primal and the primal is referred to as dual, in the case of DEA. Most people use the dual or the envelopment approach to determine the efficiency scores. Dual is less computational, as it contains $m + s$ constraints, when compared to primal which contain $n$ constraints. Envelopment model is more meaningful as it calculates the amount of slack associated with each input and output thereby providing recommendation to the management for improving the efficiency.

DEA models can be subdivided into input and output orientated models. Input oriented model aims at minimizing the input consumed by the DMUs for the same target of output levels. While output oriented models aims at maximizing the outputs produced by the DMUs for the given amount of inputs consumed. Model (4) shows the formulation of input oriented CCR model and model (5) shows the formulation of output oriented CCR model, Charnes (1978).

| CCR Input Oriented (Multiplier Approach) | CCR Output Oriented (Multiplier Approach) |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} u_r y_{0r}$<br>S.T (4)<br>$\sum_{i=1}^{m} v_i x_{0i} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \leq 0, \forall j = 1, \dots, n$<br>$u_r, v_i \geq 0$ | $Min \quad Z = \sum_{i=1}^{m} v_i x_{0i}$<br>S.T (5)<br>$\sum_{r=1}^{s} u_r y_{0r} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \leq 0, \forall j = 1, \dots, n$<br>$u_r, v_i \geq 0$ |

### *2.4.4 Banker-Charnes-Cooper (BCC) Model*

BCC model named after Banker, Charnes, and Cooper is an extension to CCR model which assumes variable returns to scale was introduced in 1984. The primary difference between the CCR and BCC models is $u_o$, free variable, in the multiplier approach and $\sum \lambda = 1$, additional constraint, in the multiplier approach. BCC model production frontier is spanned by convex hull of existing DMUs. The frontier has piecewise linear and concave characteristics which leads to variable returns to scale characterizations, Banker (1984). A free variable $u_o$ indicates decreasing returns to scale, negative free variable $u_o$ indicated increasing returns to scale, and if the free variable $u_o$ equals to zero then it indicated constant returns to scale, Cooper et al., (2007).

The relationship between the CCR and BCC models is, BCC production set is a subset of CCR production set. This means if the DMU is CCR efficient then it is definitely BCC efficient also while the converse is not true. CCR models are models are selective in allocating efficiency scores; hence CCR efficiency score always less than or equal to BCC efficiency scores.

| Primal | Dual |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} u_r y_{0r} - u_o$<br>S.T (6)<br>$\sum_{i=1}^{m} v_i x_{0i} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} - e^T u_o \leq 0$<br>$u_r, v_i \geq 0$<br>$u_o \; is \; free \; in \; sign$ | $Min \quad Z = \theta$<br>S.T (7)<br>$\theta x_{oi} - \sum_{j=1}^{n} x_{ji} \lambda_j \geq 0 \quad \forall \, i = 1, \dots, m$<br>$\sum_{j=1}^{n} y_{jr} \lambda_j \geq y_{or} \quad \forall \, r = 1, \dots, s$<br>$\sum_{j=1}^{n} \lambda_j = 1$<br>$\lambda_j \geq 0 \; \forall \, j = 1, \dots, n$ |

Similar to CCR model, BCC also possess the primal and dual models. The primary difference the CCR primal and BCC primal model is the free variable $u_o$. While the primary difference between the CCR dual and BCC dual is the additional constraint $\sum \lambda = 1$. The primary and dual models of BCC are represented as Model (6) and (7) respectively. Each primal and dual model can be sub divided into input oriented model and output oriented model. The input and output oriented BCC models based on the primal or multiplier approach are represented as Model (8) and (9) respectively.

| BCC Input Oriented (Multiplier Approach) | BCC Output Oriented (Multiplier Approach) |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} u_r y_{0r} - u_o$<br>S.T (8)<br>$\sum_{i=1}^{m} v_i x_{0i} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} - u_o \leq 0$<br>$u_r, v_i \geq 0$<br>$u_o \; is \; free \; in \; sign$ | $Min \quad Z = \sum_{i=1}^{m} v_i x_{0i} - v_o$<br>S.T (9)<br>$\sum_{r=1}^{s} u_r y_{0r} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} + v_o \leq 0$<br>$u_r, v_i \geq 0$<br>$v_o \; is \; free \; in \; sign$ |

### *2.4.5 Additive Model*

The additive model was introduced by Charnes, Clark, Cooper, and Golany in 1985. Additive models possess the advantage of combining both input and output oriented models by treating input and output slacks directly in the objective function, Charnes et al., (1985). Hence it is also known as the non-oriented model. The additive model deals with the input excesses and output shortfalls directly and can also discriminate efficient and inefficient DMUs simultaneously. This model has the same production possibility set as BCC model, based on variable returns to scale. So both these models possess the similar constraint $\sum \lambda = 1$. The multiplier approach and envelopment approach are represented as model (10) and (11) respectively. As the BCC and Additive models are based on variable returns to scale, if a DMU is BCC efficient then it is Additive efficient also.

| Additive Model (Multiplier) | Additive Model (Envelopment) |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} s_r^+ - \sum_{i=1}^{m} s_i^-$<br>S.T (10)<br>$-\sum_{j=1}^{n} x_{ij}\lambda_j + x_{io} \le 0 \quad \forall\, i = 1, \dots, m$<br>$\sum_{j=1}^{n} y_{rj}\lambda_j \le y_{ro} \quad \forall\, r = 1, \dots, s$<br>$\sum_{j=1}^{n} \lambda_j = 1$<br>$\lambda_j \ge 0 \ \forall\, j = 1, \dots, n$ | $Min \quad \sum_{i=1}^{m} v_i x_{i0} - \sum_{r=1}^{s} u_r y_{r0} + u_o$<br>S.T (11)<br>$-\sum_{i=1}^{m} v_i x_{ij} + \sum_{r=1}^{s} u_r y_{rj} - u_o \le 0$<br>$u_r \ge 1 \ \forall\, r = 1, \dots, s$<br>$v_i \ge 1 \ \forall\, i = 1, \dots, m$<br>$u_o \ is \ free \ in \ sign$ |

## 2.5 DEA Models for Special Cases

Apart from the basic models of DEA such as CCR and BCC there exists several other models in DEA which are chosen based on specific requirement. The following are few of the important models which attracted the attention of many researchers due to their specific advantages.

### *2.5.1 Non-Discretionary Variables*

Non-Discretionary variables mean those which cannot be controlled at will. These are also called as exogenously fixed variables or uncontrolled variables as per DEA terminology. The basic assumption of DEA is that all the multiple inputs and outputs considered are controllable, so inefficiencies associated with each variable can be adjusted to achieve the desired efficiency. There exists some variables which are beyond the control of management and sometime human power, but these variables need to be considered. Examples for non-discretionary variables are fertility of the farmlands, age of the store, local unemployment rate, growth of population and influence of weather.

Banker and Morey (1986a), modified CCR input oriented envelopment model to include discretionary and non-discretionary variables as shown below:

$$Min \quad \theta - \varepsilon \left( \sum_{i \epsilon I_D} s_i^- + \sum_{r=1}^{s} s_r^+ \right)$$

S.T (12)

$$\sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- = \theta x_{io} \qquad i \epsilon I_D;$$

$$\sum_{j=1}^{n} x_{ij} \lambda_j + s_i^- = x_{io} \qquad i \epsilon I_N$$

$$\sum_{j=1}^{n} y_{rj} \lambda_j - s_r^+ = y_{ro} \qquad r = 1, \ldots, s$$

$$\lambda_j \geq 0 \; j = 1, \ldots, n$$

The set of multiple inputs variables can be classified into discretionary and non-discretionary variables. Discretionary input variables are represented as $I_D$, non-discretionary input variables are represented as $I_N$. Modifications made by Banker and Morey (1986a) for the treatment of non-discretionary variables are as follows:

1. Slack variables associated with non-discretionary input variables are not introduced into the objective function. Hence inefficiencies associated with these variables do not influence the efficiency score of DMU under evaluation.
2. θ, the measure of efficiency does not control the non-discretionary input variables. It only controls the discretionary input variables.

### *2.5.2 Categorical DMUs*

All the DMUs during DEA need to be homogenous. There might be times where the DMUs are not 100% homogenous. All the DMUs in the reference will not have the same kind of advantages and disadvantages when compared to other DMUs. At the same time it is not reasonable to compare the disadvantageous DMUs with advantageous DMUs. DEA structure with hierarchical category is required, such that DMUs with similar advantages are compared among themselves and also with more disadvantages (worst) DMUs. Whereas more disadvantaged DMUs are compared among themselves. If these categories are not comparable then separate analysis is required.

For example consider a set of hospitals. These hospitals which are based on population distribution around them can be classified into two categories, advantageous and disadvantageous. Hospitals in disadvantageous category are compared among themselves and not with the advantageous; whereas hospitals in advantageous category are compared among themselves as well as with the hospitals in disadvantageous category.

Banker and Morey (1986b), modified version of the CCR input oriented envelopment model including categorical DMUs is shown below:

$$\text{Min } \theta$$

$$\text{S.T} \qquad\qquad\qquad\qquad\qquad\qquad (13)$$

$$\sum_{j \in \bigcup_{f=1}^{K} K_f} x_{ij}\lambda_j + s_i^- = \theta x_{io} \qquad i = 1, \ldots, m$$

$$\sum_{j \in \bigcup_{f=1}^{K} K_f} y_{ij}\lambda_j - s_r^+ = y_{io} \qquad r = 1, \ldots, s$$

$$\lambda_j \geq 0 \quad j = 1, \ldots, n$$

Determine the total number of possible categories such that each DMU can be assigned to a particular category. Let’s say there exist ‘L’ categories (1 ≤ f ≤ L), such that ‘1’ represents the lowest level and ‘L’ represent the highest level. DMUs in each category from (1 ≤ f ≤ L) are represented as $K_1$, $K_2$, …, $K_L$. Each DMU can be assigned to only particular category and every DMU in the set should be assigned to some category. All DMU which belong to category 1 are evaluated with respect to units in $K_1$, all DMUs in category 2 are evaluated with respect to units in $K_1$ U $K_2$.

### *2.5.3 Weight Based Models*

DEA calculates weights for multiple inputs and outputs to evaluate relative efficiency by maximizing the ratio of weighted sum of outputs to inputs. Restrictions on weights are non negativity, and efficiency of DMU should be less than unity. The flexibility of weights in DEA is considered to be both advantageous and disadvantageous. The weights associated are sometimes beyond the scope of explanation. This led to the development of the weight restriction models, which confines selection of weights to a finite limit.

Weight restricted models are application oriented. Formulation of initial weight restricting model happened when CCR model failed to choose the best site for locating a high energy physics lab. Thompson (1986) restricted input and output weights using Assurance Region model (AR). This led to a new era of weight restricted models in DEA.

Assurance Region approach developed by Thompson (1986) imposes constraints based on ratio of relative magnitude of weights for input and output variables. The two ways in determining Assurance Region bounds are, using Analytical Hierarchy Process (AHP) to obtain expert suggestions for setting the bounds, and the second one is using economic information of price/unit cost. Dyson (1988) presents a rationale for limiting weights in DEA models with single input case, and regression analysis for determining the lower bounds. This cannot be applied as a general model for multiple inputs and outputs. Cone Ratio method developed by Charnes (1990) is a more general approach than Assurance Region approach. Convex cones are used to measure the efficiency of DMUs. The feasible regions of weights are reduced to be a polyhedral convex cone by using directional vectors carrying conditions specified by the decision maker. Roll (1991) assuming the extreme case, when no flexibility is allowed, determined the concept of Common Set of Weights (CSW), which is a usual approach in efficiency analysis. The common set of weights is determined by taking average of the upper and lower bound weights from unbounded DEA analysis. Roll (1993) proposes that the weight bounds can be chosen by carefully observing the resultant weight matrix of the unbounded DEA model. Possible consideration would be eliminating zero weights and finding average weights for each factor across all DMUs.

### *2.5.4 Super Efficiency Model*

Super efficiency model is introduced with the objective of providing tie-breaking method among the efficient DMUs and effective procedure for ranking DMUs. The process of excluding the DMU under evaluation from the solution set results in a new set of efficiency scores. Ranking of the super efficiency model is based on the new solution set. The model obtained by excluding the data of the decision making unit from the reference set of the envelopment model to calculate the efficiency score is called as Super Efficiency Model. This model was proposed by Andersen and Petersen (1993). This model takes the form of CCR and avoids the convexity constraint condition of BCC model.

$$Min \quad Z = \theta$$

$$\text{S.T} \qquad\qquad\qquad\qquad\qquad\qquad (14)$$

$$\theta x_{oi} - \sum_{j=1,\neq 0}^{n} x_{ji}\lambda_j \geq 0, \ \forall\, i = 1, \ldots, m$$

$$\sum_{j=1,\neq 0}^{n} y_{jr}\lambda_j \geq y_{or}, \forall\, r = 1, \ldots, s$$

$$\lambda_j \geq 0 \ \forall\, j = 1, \ldots, n$$

## 2.6 Strengths and Weaknesses

Data Envelopment Analysis is more advantageous when compared to other efficiency measurement techniques.

- DEA is a Total Factor Productivity (TFP) approach since it has the capability to handle multiple inputs and outputs, unlike Ratio Analysis technique which is limited to single input and output.
- DEA is a non-parametric approach for which relationship between inputs and outputs need not be defined, unlike a parametric approach where accuracy of defining the relationship between inputs and outputs could influence the results.
- Units Invariance property of DEA models implies that final results are independent of unit measurement of inputs and outputs, provided the units are same for every DMU, Cooper et al., (2007).
- Unlike the regression analysis models providing focus on mean values of the group, DEA measures the amount of efficiency and inefficiency associated with each individual unit.
- DEA also calculates the required projections for transferring inefficient units to be efficient and provides information about the benchmarks used.

- Positivity condition (e >0) of DEA multipliers provides flexibility for each DMU to be evaluated as the best, Cooper et al., (2004).
- Unlike the fixed weight models DEA is more advantageous with flexibility in choosing variable weights to represent each DMU in its best form.

Like any other method DEA also has few weaknesses.

- DEA is only capable of measuring relative technical efficiency and it cannot measure absolute efficiency. This implies that 100% technical efficient units are best among the peers but may not be 100% absolute efficient.
- DEA is a frontier technique and is extremely sensitive to quality of the data and outliers present in the data which can greatly influence the estimation of frontier.
- DEA efficiency results are dependent on data provided. Any addition of inputs and/or outputs and addition of new DMUs can influence existing efficiency scores.
- The rule of thumb by Banker (1989) suggests that number of DMUs should be greater than or equal to the maximum of either the product value of inputs and outputs or thrice the sum of inputs and outputs. This limits the application of DEA to smaller sets.

$$n \geq \max\{m \times s, 3(m+s)\}$$

$$n = Number\ of\ DMUs, \qquad m = Number\ of\ inputs, \qquad s = Number\ of\ outptus$$

## 2.7 Healthcare Efficiency Measurements using DEA

Tracing the subject of the application of DEA in healthcare, Nunamaker (1983) was the pioneer who made a comparison between the cost saving estimates per patient day and efficiency scores of DEA models. 17 fairly homogenous hospitals in Wisconsin were selected for measuring the efficiency of nursing services over a two year period starting from 1978. Input measure is aggregation of all routine costs associated with inpatients and output measure is aggregation of patient days. The results represented fundamental differences between the two methodologies and differences in efficiency scores for different combinations of inputs and outputs. Nunamaker provided a glance of DEA application in healthcare and Sherman (1984) provided more strength through his satisfactory research findings. Sherman suggested DEA as a promising tool to evaluate hospital efficiency when compared to other approaches such as ratio analysis, and econometric regression analysis. When DEA was applied to a group of teaching hospitals in Massachusetts; it was found that DEA provided better insight into location and nature of hospital inefficiencies, and identified two inefficient clinics that would not be identified with other efficiency approaches. Banker (1986) added strength to application of DEA in healthcare by comparing the DEA results with translog cost functions. Data of 114 hospitals in North Carolina for the fiscal year of 1978 was examined using four inputs and three outputs. Inferences suggested that efficiency scores of DEA were highly correlated to the actual capacity utilization estimated by hospitals than compared to results of translog estimates. DEA results estimate diverse set of behaviors under increasing and decreasing (variable returns to scale) production functions whereas translog functions estimate using constant returns to scale.

The capability of DEA to handle multiple inputs and outputs, non-parametric nature, focus on each individual unit, and the ability to measure the efficiency score under variable returns to scale production functions provided break through for application of DEA to measure efficiency of healthcare units. The saga of successful application of DEA in healthcare continued and expanded to wider areas within a short period of time. Following is the list of few applications of DEA in wider areas of healthcare. Ozcan (1998) identified physician benchmarking in treatment of Otitis media using CCR model. Siddharthan (2000) used DEA to measure the relative technical efficiency of 164 Health Maintenance Organizations (HMOs) in United States. Healthcare utilization was measured using inpatient days, number of outpatients, emergency room visits as input measures and output measures are number of commercial, Medicaid, and Medicare life's covered in each plan. Nathanson (2003) used DEA to identify survival chances of Neurotrauma patients at an early stage during their stay in the Intensive Care Unit (ICU). Variables that influence death of the patient during his stay in ICU are considered as inputs. The efficiency score of the DEA results indicate the survival chance of the patients. Higher efficiency score indicates better chance of recovery. Nathanson compared the performance of DEA results against regression models. DEA results are more satisfactory as each patient efficiency level can be identified than focusing on mean values of the group. Basson (2006) performed Data Envelopment Analysis to evaluate operating room efficiency across 23 Veteran Health Administration systems. The results conclude that DEA is capable of providing information more specifically about efficiency and inefficiency for each unit when compared to single ratio methods. Mukherjee (2010) analyzed the efficiency of Local Health Departments (LHD) operating in U.S based on 2005 data using DEA.

Taking a turn to consider the application of DEA in healthcare in geographic locations outside United States, it has been applied in many other countries to measure the efficiency of healthcare units. Garavaglia (2011) evaluated efficiency and quality care of 40 nursing homes in Italy over a three year period starting from 2005. A two stage DEA analysis procedure was used. Blank (2009) used DEA to identify the productivity of 69 Dutch hospitals based on data for the year 2000. This particular research work also employed a two stage process; efficiency scores are determined based on DEA models in the first stage and in the second stage bootstrapping techniques were used to identify the factors influencing the costs and inefficiencies in hospitals. Puenpatom (2008) used a two stage DEA approach to measure the efficiency of public hospitals in Thailand during the transition phase of implementing new health coverage. The research was able to identify increase in efficiency of larger public hospitals during the transition period.

Hofmarcher (2002) preferred DEA models to measure the efficiency of Austrian hospitals, for the years 1994 to 1996, over regression models and fixed effects model. Helmig (2001) used DEA to measure the efficiency of public, welfare, and private hospitals in Germany for the years 1991 to 1996. Results suggest that overall efficiency of the hospitals over the period of time had increased. The research work also drew inferences about the influence of ownership on efficiency scores, as the results suggested that public and welfare hospitals are relatively more efficient than private hospitals. Bjorkgren (2001) used DEA to identify the nursing care efficiency of 64 long term care units in Finland based on four inputs and one output. In the second stage, statistical significance test were carried out to compare the efficiency scores.

Hollingsworth (2001) expressed DEA as a potential tool to evaluate the efficiency of 49 neonatal care services in U.K over regression analysis. The research work used the dataset published by O’Neill, who determined the average costs using regression analysis. Results of DEA models suggested that there exists more potential scope for savings when compared to regression analysis, and significant technical inefficiencies due to economic returns to scale. Jacobs (2001) examined the efficiency of 232 U.K hospitals based on seven different measures to study the consistency and robustness of efficiency scores using Data Envelopment Analysis and Stochastic Frontier Analysis. The results suggested that both these methods possessed strengths and weakness, taking into account the random noise in data. If there was no random noise in the data both these methods provide consistent results.

Hollingsworth (1999) reviewed application of non-parametric approach in measuring efficiency of healthcare units with major focus on Data Envelopment Analysis. This work reviewed a total 91 published papers starting with Nunamaker (1983) to 1999. The statistics published by Hollingsworth (2003) shows the growth of DEA studies in healthcare over a period of time, percentage of studies using DEA methodology. The study finds that DEA tops the list when compared to other efficiency measurement studies and healthcare is the primary area.

Hollingsworth (2003) reviewed 188 published papers identifying the application of parametric and non-parametric approaches in measuring efficiencies of healthcare units which is an extension to Hollingsworth 1999. The statistics suggested that almost 50% of the literature used DEA alone to identify the efficiency scores. 12% of the studies used Stochastic Frontier Analysis (SFA) and other parametric approaches; this leaves 88% of the studies which used DEA alone or combined with some other methods to measure the efficiency of healthcare units. These statistics indicated the significance of DEA for healthcare efficiency measurement, and for much more details refer to Hollingsworth 1999, 2003.

Based on the Bibliography work done by Becker (2010) starting from the inception of DEA in 1978, the list of healthcare related journals and the number of published research articles in the field of DEA are presented in Appendix A.

# Chapter 3 - PREPARING THE DATA

This chapter introduces the mission, objectives, and programs of Kansas Association for the Medically Underserved (KAMU). Overview of data provided by the KAMU for the determination of the productivity will be presented in this chapter. It also presents the list of member clinics whose productivity needs to be determined. Summary of the problems associated with the data and the measures need to be taken, to prepare the data for Data Envelopment Analysis will be addressed in this chapter.

Important inputs and outputs required for the analysis will be identified and their significance will be presented in this chapter. Some of the issues such as classification of inputs and outputs, correlation between the variables, scaling the data, selection of DEA models, and missing data will be addressed. The measures to supplement them will also be addressed by providing insight to valuable literature works. This chapter also presents the literature review on DEA studies with missing data. The literature review addresses the methods to handle the missing data during Data Envelopment Analysis.

This chapter is structured as follows. Section 3.1, introduces the history of KAMU as well as its mission and aims. Section 3.2, presents the overview of the data provided by the KAMU. Section 3.3, describes the significance of important variables from the view point of the KAMU clinics. Section 3.4, classifies the identified list of variables as inputs and outputs. It also presents the literature in this aspect to classify inputs and outputs more effectively. Section 3.5, provides the guidelines to choose the specific DEA model required. Section 3.6, highlights the influence of correlation on DEA outcomes, while section 3.7 highlights the influence of normalization on DEA outcomes. Section 3.8, presents the issue of missing data, core aspect of this thesis. Then section 3.9, digs into the literature works on DEA with missing data. Section 3.10, introduces the software used to execute the various DEA models as are part of this thesis.

## 3.1 Introduction to KAMU

Kansas Association for the Medically Underserved was founded in 1989 and incorporated as nonprofit organization in 1990. In 1991 KAMU was recognized as the Primary Care Association (PCA) of Kansas. KAMUs mission is to “support and strengthen its member organizations through advocacy, education and communication”. It provides advocacy on behalf of Federally Funded or Locally Funded Community Health Centers in Kansas, education by providing training and technical assistance, and communication among the clinics for sharing beast practices and to improve the knowledge. Currently 42 organizations are member clinics of KAMU. Members include both public and private non-profit organizations. The main aim of these clinics is to “deliver primary health care services regardless of an individual's ability to pay”. Members range from Federally Qualified Community Health Centers (FQHCs) to the local county health departments. Membership is open to all organizations which can meet KAMU membership criteria, support it missions, and practice the aim.

Currently there are 42 KAMU member clinics which include 14 Federally Qualified Health Centers (FQHC), 1 FQHC Look-Alike, 26 Primary Care Clinics; other member does not provide direct care but supports KAMU mission. The data provided by KAMU in 2008 include 41 clinics, of which 19 Federally Qualified Health Centers (FQHC), 14 Primary Care Clinics, 7 Free Clinics, and 1 Voucher Program. The list of all the current 42 member clinics under KAMU is shown in Appendix B.

The association accomplishes its mission through a wide range of programs and activities that can be grouped into seven core functions:

1) **Maintaining and Strengthening the State's Safety Net Primary Care Clinics:** KAMU provides training and technical assistance to health centers and primary care safety net clinics, as well as targeted assistance to new clinics and organizations with leadership or other significant changes.
2) **Surveillance**: KAMU monitors state regulatory, administrative and legislative activities that affect the need for and availability of primary care services for the underserved. Facilitating activities to positively influence and impact outcomes that affect the underserved.
3) **Growth Assistance:** KAMU assists existing organizations and communities to expand primary health care services for underserved populations consistent with their need.
4) **Workforce**: KAMU helps member organizations with the development of recruitment and retention plans, partners with National Health Service Corps and State Loan Repayment offices, and works with academic medical, dental and other health profession schools to promote the placement of student interns in safety net clinics.

5) **Liaison/Collaboration/Partnering:** KAMU works collaboratively with local, state and federal officials and organizations involved in health policy.
6) **Clinical Quality:** KAMU provides technical assistance to health centers and primary care safety net clinics on care management and clinical quality activities and programs, as well as emergency preparedness planning.
7) **Leveraging and Enhancing Revenues:** KAMU works with private and public stakeholders to increase resources for operations and/or capital improvements to improve Kansans' access to quality primary health care services.

KAMU offers a variety of programs to support member clinics in their work and to expand health care for the underserved in Kansas.

- Clinical Programs provide access to resources and information that enhance patient care and clinical proficiency. Additionally, they provide opportunities and support that connect clinicians working in safety net clinics across the state.
- Community Development is aimed at growing, strengthening and sustaining the primary health care safety net in Kansas.
- Workforce Development programs help recruits and train professionals to work in the state's primary care safety net clinics.
- Operational and Financial services are targeted to help KAMU members develop strategies to provide financial stability and increase operating efficiencies.

## 3.2 Introduction to KAMU Data

The 41 clinics data provided by KAMU for Data Envelopment Analysis is the outcome of Clinic Reporting Tool (CRT) used by these clinics. The information collected by this tool is developed based on the requirements of the KAMU and KDHE, Kansas Department of Health and Environment. The primary objective of this tool is to reduce the multiple reporting burdens of the clinics. The clinics need to submit this data once in a year to KDHE.

The data collected by the CRT is provided in an Excel File. Data in the spreadsheets is broadly classified into the following sections: List of Clinics, Expenses, Revenue, Staffing, Diagnosis, Patient Visits, and the List of Services offered by each clinic. Each section has data recorded for large number of variables, and the total number of variables in the dataset is 225.

Major issue associated with the data provided is, a large amount of it is missing. Certainly all the 225 attributes are not required for the Data Envelopment Analysis. The relationship between the number of attributes required for the analysis and the number of the DMUs is presented below. This illustrates the requirement for total number of inputs and outputs.

1. Number of DMUs should be equal or greater than the product of inputs and outputs

or

2. Number of DMUs should be equal or greater than 3 time the sum of inputs and outputs

$$n \geq Max\ \{3(m+s), m * s\}$$

$$Where\ n - \#\ of\ DMUs, m - \#\ of\ inputs, s - \#\ of\ outputs$$

Based on this relationship the total number of inputs and outputs should not be more than 13 attributes, one third of 41 DMUs. There is no particular restriction for choosing the number of inputs and outputs among the 13 variables. The list of important inputs and outputs required for the analysis based on the availability of the data from each section is presented in the Table 3.1.

**Table 3-1: List of Important Inputs and Outputs**

| Expenses | Revenue | Staffing | Patient Visits |
|---|---|---|---|
| Medical Staff | Medicaid Charged | Nurse Practitioner FTE | Total Users |
| Lab X-Ray | Medicaid Collected | Nurse Practitioner Enc | Uninsured Users |
| Other Medical | Self Pay Charged | Nurses FTE | Total Visits |
| Facility | Self Pay Collected | Administration FTE | |
| Administration | State PC Collected | Patient Support FTE | |

Examination of the KAMU data, regarding the missing values, confirms that there exist few clinics (DMUs) with high percentage of missing values. Elimination of such clinics from the analysis could reduce the effort to estimate missing values. There exist three such clinics in the dataset which does not have data for most of the variables presented in Table 3.1. There exist one more DMU which does not have any input data required for the analysis. Estimating all the inputs might affect DEA results. Excluding these four DMUs from the dataset, the total number of DMUs reduces from 41 to 37. The percentage of the data available for each of the variables presented in the Table 3.1 is shown below in Table 3.2

**Table 3-2: Data Availability for Each Variable**

| Key No | Variables | % Data Available |
|---|---|---|
| V1 | Medical Staff Expenses | 89.19 |
| V2 | Lab X-Ray Expenses | 70.27 |
| V3 | Other Medical Expenses | 72.97 |
| V4 | Facility Expenses | 81.08 |
| V5 | Administration Expenses | 86.49 |
| V6 | Medicaid Charged | 67.57 |
| V7 | Medicaid Collected | 67.57 |
| V8 | Self Pay Charged | 86.49 |
| V9 | Self Pay Collected | 91.89 |
| V10 | State PC Collected | 75.68 |
| V11 | Nurse Practitioner FTE | 72.97 |
| V12 | Nurse Practitioner Enc | 62.16 |
| V13 | Nurses FTE | 78.38 |
| V14 | Administration FTE | 83.78 |
| V15 | Patient Support FTE | 70.27 |
| V16 | Total Users | 100.00 |
| V17 | Uninsured Users | 89.19 |
| V18 | Total Visits | 100.00 |

Identifying the potential clinics with high percentage of available data and critical variables required for the analysis, the dataset is reduced from 41 clinics to 37 clinics and the number of variables is reduced from 225 to 18. Thus based on the primary analysis the dataset is reduced and simplified for the Data Envelopment Analysis.

In order to evaluate the effectiveness of missing data estimation methods in the later chapters we need to identify a subset with complete data from the given dataset. Scrutinizing the given dataset we identified 22 clinics with 7 variables which possess complete data. This dataset will be introduced in the later chapters to evaluate the effectiveness of missing data estimation methods proposed in this thesis.

## 3.3 Description of Variables

The previous identifies the list of important variables, and explains how the large dataset is simplified to perform the Data Envelopment Analysis. This section presents the description for the list of important variables identified in the previous section based on the user manual of Clinic Reporting Tool.

**Medical Staff Expenses:**

Medical Staff Expenses includes all staff costs, including salaries and fringe benefits for personnel supported directly or under contract for medical care staff, except lab and x-ray staff.

**Lab X-Ray Expenses:**

Lab X-Ray Expenses includes all costs for lab and x-ray, including salaries and fringe benefits for personnel supported directly or under contract, for lab and x-ray staff. It also includes all other direct costs, but not limited to, supplies, equipment depreciation, related travel, contracted or voucher lab and x-ray services, etc.

**Other Medical Expenses:**

Other Medical Expenses includes all other direct costs for medical care including, but not limited to, supplies, equipment depreciation, related travel, etc.

**Facility Expenses:**

Facility Expenses include rent or depreciation, interest payments, utilities, security, grounds keeping, maintenance, janitorial services, and all other related costs.

**Administration Expenses:**

Administrative costs include the cost of all corporate administrative staff, billing and collections staff, medical records and intake staff, and the costs associated with them including, but not limited to, supplies, equipment depreciation, travel, etc.

In addition other corporate costs example purchase of insurance, audits, Board of Director's costs, etc. The cost of all patient support services example medical records and intake are also included.

**Staffing:**

All paid staff should are considered as full-time equivalents (FTEs). A person who works 20 hours per week (i.e., 50% time) is reported as "0.5 FTE." Positions with less than a 40-hour base, especially clinicians, should be calculated on whatever they have as a base for that position. Similarly, an employee who works four months out of the year would be reported as "0.33 FTE".

All staff time should be allocated by function among the major service categories listed. For example, a full-time nurse who works solely in provision of direct medical services would be counted as 1.0 FTE under "Nurses". If that nurse provided case management services for 10 hours per week, and provided medical care services for the other 30 hours per week, time would be allocated 0.25 FTE to "Case Managers" and 0.75 FTE to "Nurses".

**Administration FTE:**

Executive director, medical director, physicians or nurses with corporate (not clinical) administrative responsibilities, secretaries, fiscal and billing personnel, all other support staff and staff with administrative responsibilities. Patient support staff is also a part of the administration staff whose primary responsibilities are patient intake and their medical records.

**Medicaid:**

Clinics should report the revenue generated through all services paid for by Medicaid, regardless of whether they are paid directly or through a fiscal intermediary or any Health Maintenance Organization (HMO).

**Self Pay:**

Clinics should report as the revenue generated through all services and charges where the responsible party is the patient, including charges for indigent care programs. This also includes the amount received for all the uncovered services and individual users without insurance.

**State Primary Care (PC) Collected:**

All the revenue received State and from the State Primary Care Grant should be reported as State PC Collected.

## 3.4 Classification of Inputs and Outputs

The basic Data Envelopment Analysis model assumes that the status of each input and output is known in prior. However this is not possible under all cases, as the complexity of the analysis increases it becomes challenging to classify inputs and outputs. There are certain variables which can act as either input or output such variables are called as flexible measures. Let’s consider an example provided by Cook and Zhu (2007) on how interns in hospitals should be considered for determining the productivity. Such a factor clearly constitutes an output measure for a hospital, being one form of training provided by the organization, but at the same time it is an important component of the hospital’s total staff, hence is an input. Generally the decision on such flexible measures is left to the analyst. Still there exist methods on how to classify such measures as either input or output for more accurate analysis.

Cook et al., (2006) presents a methodology for dealing with factors which can simultaneously act as input and output. Flexible factors can be classified into three groups such as input, output, and equilibrium. These are tested under constant, increasing, and decreasing returns to scale, permitting them for allocations. Cook and Zhu (2007) proposed a new model for classifying a measure into an input or output by introducing a large positive number into the model. Later Toloo (2009) identified that such a measure can result in inaccurate efficiency score. A new model is developed by Hatefi et. al., (2009) to classify inputs and outputs based on Cobb-Douglas production function.

Examination of the list of important variables presented in the Table 3.1 reveals that there exist no such flexible measures in this analysis. The procedure for classifying the inputs and outputs is completely based on the expert opinion and meticulous understanding of the nature of the variable. The outcome suggests that Expenses and Staffing variables should be treated as inputs, whereas Revenue and Patient Visit variables should be treated as outputs. The criterion behind this classification is also based on the basic concept of inputs and outputs. Inputs can be considered as resources utilized for the achieving the goals and objectives of the system. The response of the system to the inputs consumed can be defined as output to the system. Table 3.3 presents the classified inputs and outputs.

**Table 3-3: List of Inputs and Outputs**

| Inputs | Outputs |
|---|---|
| Medical Staff Expenses | Medicaid Charged |
| Lab X-Ray Expenses | Medicaid Collected |
| Other Medical Expenses | Self Pay Charged |
| Facility Expenses | Self Pay Collected |
| Administration Expenses | State PC Collected |
| Nurse Practitioner FTE | Total Users |
| Nurse Practitioner Enc | Uninsured Users |
| Nurses FTE | Total Visits |
| Administration FTE | |
| Patient Support FTE | |

## 3.5 Selection of DEA Models

This section highlights the significance of choosing the appropriate models to perform the Data Envelopment Analysis more effectively.

In the second chapter we had discussed that variables can be classified as Discretionary and Non-Discretionary variables. Discretionary variables can be controlled by the management and Non-Discretionary variables are beyond the control of the management. The inputs, Expenses and Staffing, fall under the category of the Discretionary variables since they are controllable. The outputs, Revenue and Patient Visits, are not under the control of management hence they fall under the Non-Discretionary variables category. Application of the Non-Discretionary models, discussed in the Chapter 2, will provide a better set of efficiency scores for the KAMU clinics. This can also be substituted by choosing appropriate input or output oriented basic DEA models.

In chapter 2, we also come across the input oriented model, and output oriented model of DEA. We also discussed the significance of these models and when they need to be applied. Input oriented model aims at reducing the amount of inputs for the same value of outputs. Whereas output oriented model aims at increasing the amount of outputs for the same value of inputs. Based on the nature of these definitions and since only the inputs, Expenses and Staffing, are discretionary variables, input oriented models are considered to be as the best choice. Thus the projection scores obtained by the DEA can be implemented effectively for improving the productivity of the inefficient clinics. Table 3.4 shows the basic input oriented DEA models.

**Table 3-4: Input Oriented Models**

| CCR Input Oriented Model | BCC Input Oriented Model |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} u_r y_{0r}$<br>S.T (1)<br>$\sum_{i=1}^{m} v_i x_{0i} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} \leq 0, \forall j = 1, \dots, n$<br>$u_r, v_i \geq 0$ | $Max \quad Z = \sum_{r=1}^{s} u_r y_{0r} - u_o$<br>S.T (2)<br>$\sum_{i=1}^{m} v_i x_{0i} = 1$<br>$-\sum_{i=1}^{m} v_i x_{ji} + \sum_{r=1}^{s} u_r y_{jr} - u_o \leq 0, \forall j = 1, \dots, n$<br>$u_r, v_i \geq 0, u_o \; is \; free \; in \; sign$ |

## 3.6 Influence of Correlation

This particular section highlights the significance for checking the correlation between the inputs and outputs prior to the analysis. The presence of high correlation between the variables does not affect the efficiency scores or the final outcomes. Identification of such variables could save the computational time, since one of the two variables can be excluded from the analysis. This can be explained with the help of an example.

Consider a dataset of 10 DMUs with single input and single output shown in the Table 3.5. The calculated efficiency scores using CCR and BCC input oriented models are also presented in the same Table. Now let's add new input variable to the dataset, such that the second input is obtained by multiplying a constant value to the first input. This shows that the correlation value between the first and second input is one. The calculated efficiency scores using CCR and BCC input oriented models after adding the additional variable are also presented in the same Table. Comparison of the efficiency scores before and after the additional variable shows that there exists no difference between them.

**Table 3-5: Influence of Correlation**

| Dataset | | | Efficiency Score before | | Additional Input Variable | Efficiency Score after | |
|---|---|---|---|---|---|---|---|
| DMU | (I) I1 | (O) O1 | CCR-I | BCC-I | | CCR-I | BCC-I |
| E01 | 25 | 19 | 0.827 | 1.000 | 37.5 | 0.827 | 1.000 |
| E02 | 28 | 20 | 0.777 | 0.921 | 42 | 0.777 | 0.921 |
| E03 | 40 | 27 | 0.735 | 0.785 | 60 | 0.735 | 0.785 |
| E04 | 27 | 16 | 0.645 | 0.926 | 40.5 | 0.645 | 0.926 |
| E05 | 33 | 29 | 0.956 | 1.000 | 49.5 | 0.956 | 1.000 |
| E06 | 34 | 29 | 0.928 | 0.971 | 51 | 0.928 | 0.971 |
| E07 | 30 | 24 | 0.871 | 0.967 | 45 | 0.871 | 0.967 |
| E08 | 28 | 16 | 0.622 | 0.893 | 42 | 0.622 | 0.893 |
| E09 | 26 | 20 | 0.837 | 0.992 | 39 | 0.837 | 0.992 |
| E10 | 37 | 34 | 1.000 | 1.000 | 55.5 | 1.000 | 1.000 |

This can also be explained mathematically. The concept of adding a new input or output variable to the dataset implies adding a new constraint to the linear programming structure of CCR and BCC models. The new constraint added is linearly dependent to one of the existing constraints. The presence of similar linear constraints in the linear programming problem does not influence the results. Hence addition of such new input or output variables to the data does not change DEA efficiency scores.

From the above example the need for calculating correlation between the input and output variables can be identified. Sometime the correlation between the inputs and outputs can be natural and sometimes it is accidental. Table 3.6 provides the correlation between the inputs and outputs identified based on the Key No provided in Table 3.2.

**Table 3-6: Correlation between the different inputs and outputs of KAMU data**

| | V1 | V2 | V3 | V4 | V5 | V6 | V7 | V8 | V9 | V10 | V11 | V12 | V13 | V14 | V15 | V16 | V17 | V18 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| V1 | 1.00 | 0.65 | 0.60 | 0.72 | 0.73 | 0.45 | 0.72 | 0.83 | 0.76 | 0.62 | 0.71 | 0.65 | 0.85 | 0.82 | 0.59 | 0.78 | 0.67 | 0.85 |
| V2 | 0.65 | 1.00 | 0.45 | 0.82 | 0.65 | 0.49 | 0.59 | 0.78 | 0.59 | 0.48 | 0.68 | 0.68 | 0.46 | 0.59 | 0.53 | 0.71 | 0.67 | 0.83 |
| V3 | 0.60 | 0.45 | 1.00 | 0.32 | 0.17 | 0.23 | 0.31 | 0.41 | 0.49 | 0.31 | 0.33 | 0.36 | 0.34 | 0.37 | -0.05 | 0.53 | 0.46 | 0.60 |
| V4 | 0.72 | 0.82 | 0.32 | 1.00 | 0.79 | 0.53 | 0.66 | 0.80 | 0.62 | 0.33 | 0.66 | 0.54 | 0.56 | 0.72 | 0.69 | 0.73 | 0.66 | 0.83 |
| V5 | 0.73 | 0.65 | 0.17 | 0.79 | 1.00 | 0.52 | 0.81 | 0.83 | 0.61 | 0.54 | 0.69 | 0.68 | 0.55 | 0.87 | 0.77 | 0.74 | 0.63 | 0.80 |
| V6 | 0.45 | 0.49 | 0.23 | 0.53 | 0.52 | 1.00 | **0.90** | 0.48 | 0.30 | 0.16 | 0.48 | 0.41 | 0.47 | 0.57 | 0.44 | 0.52 | 0.21 | 0.61 |
| V7 | 0.72 | 0.59 | 0.31 | 0.66 | 0.81 | **0.90** | 1.00 | 0.79 | 0.41 | 0.37 | 0.53 | 0.49 | 0.66 | 0.76 | 0.60 | 0.73 | 0.57 | 0.80 |
| V8 | 0.83 | 0.78 | 0.41 | 0.80 | 0.83 | 0.48 | 0.79 | 1.00 | 0.79 | 0.58 | 0.79 | 0.76 | 0.61 | 0.84 | 0.57 | **0.91** | 0.88 | **0.91** |
| V9 | 0.76 | 0.59 | 0.49 | 0.62 | 0.61 | 0.30 | 0.41 | 0.79 | 1.00 | 0.59 | 0.87 | 0.83 | 0.47 | 0.72 | 0.47 | 0.77 | 0.69 | 0.76 |
| V10 | 0.62 | 0.48 | 0.31 | 0.33 | 0.54 | 0.16 | 0.37 | 0.58 | 0.59 | 1.00 | 0.67 | 0.84 | 0.48 | 0.52 | 0.37 | 0.50 | 0.44 | 0.55 |
| V11 | 0.71 | 0.68 | 0.33 | 0.66 | 0.69 | 0.48 | 0.53 | 0.79 | 0.87 | 0.67 | 1.00 | **0.93** | 0.45 | 0.71 | 0.61 | 0.77 | 0.65 | 0.80 |
| V12 | 0.65 | 0.68 | 0.36 | 0.54 | 0.68 | 0.41 | 0.49 | 0.76 | 0.83 | 0.84 | **0.93** | 1.00 | 0.41 | 0.68 | 0.48 | 0.75 | 0.64 | 0.76 |
| V13 | 0.85 | 0.46 | 0.34 | 0.56 | 0.55 | 0.47 | 0.66 | 0.61 | 0.47 | 0.48 | 0.45 | 0.41 | 1.00 | 0.69 | 0.57 | 0.61 | 0.46 | 0.59 |
| V14 | 0.82 | 0.59 | 0.37 | 0.72 | 0.87 | 0.57 | 0.76 | 0.84 | 0.72 | 0.52 | 0.71 | 0.68 | 0.69 | 1.00 | 0.61 | 0.81 | 0.68 | 0.82 |
| V15 | 0.59 | 0.53 | -0.05 | 0.69 | 0.77 | 0.44 | 0.60 | 0.57 | 0.47 | 0.37 | 0.61 | 0.48 | 0.57 | 0.61 | 1.00 | 0.51 | 0.37 | 0.62 |
| V16 | 0.78 | 0.71 | 0.53 | 0.73 | 0.74 | 0.52 | 0.73 | **0.91** | 0.77 | 0.50 | 0.77 | 0.75 | 0.61 | 0.81 | 0.51 | 1.00 | **0.91** | **0.92** |
| V17 | 0.67 | 0.67 | 0.46 | 0.66 | 0.63 | 0.21 | 0.57 | 0.88 | 0.69 | 0.44 | 0.65 | 0.64 | 0.46 | 0.68 | 0.37 | **0.91** | 1.00 | 0.81 |
| V18 | 0.85 | 0.83 | 0.60 | 0.83 | 0.80 | 0.61 | 0.80 | **0.91** | 0.76 | 0.55 | 0.80 | 0.76 | 0.59 | 0.82 | 0.62 | **0.92** | 0.81 | 1.00 |

If the correlation value between the variables is greater than 0.90 then one out of the two variables is dropped from the analysis. The correlation between Medicaid Charged (V6) and Medicaid Collected (V7) is greater than 0.90. Hence one out of them can be dropped from the analysis. Medicaid collected includes the bad debts, money failed to collect from patients due to economical conditions or other issues. These bad debts varies from clinic to clinic, hence it would be better be to drop variable Medicaid Collected (V7) from the analysis.

Similarly based on the correlation between the variables and due to availability of the data some of the variables are dropped from the analysis. The final set of input and output variables identified for the analysis are shown in the Table 3.7.

Table 3-7: Final List of Inputs and Outputs

| Key No | Input Variables | Key No | Output Variables |
|---|---|---|---|
| I1 | Medical Staff Expenses | O1 | Medicaid Charged |
| I2 | Facility Expenses | O2 | Self Pay Charged |
| I3 | Nurses F.T.E | O3 | Total Users |
| I4 | Administration F.T.E | | |

## 3.7 Normalization

DEA efficiency score is defined as the ratio of weighted sum of outputs to weighted sum of inputs. This signifies the importance of weights in DEA. Input and Output weights are outcomes of the linear programming methodology. DEA provides high value of weight to the most favorable input or output to bestow the target DMU with best efficiency scores. Assignment of these weights is influenced by the variations in the data.

There exist large variations in the magnitudes of the inputs and outputs of the KAMU data. The expenditure and revenue values are in Millions ($), where as the Full Time Employment (FTE) is in Tens. Thus there exist large variations in the data. The variation in the magnitude of the data reflects in the magnitude of weights. Hence normalization is the better way to obtain the similar magnitude among the data. The efficiency scores of the DMUs will not be affected by the normalization process as the DEA models are unit invariant and independent scale transformations for the input and output variables are allowed, Cooper et al., (2007).

## 3.8 Missing Data

The last but the most important obsession that need to be addressed for preparing the data is missing values. Traditionally Data Envelopment Analysis requires availability of complete data for each input and output to perform the analysis, with the data assumed to be positive for all DMUs. Most of the practical applications do not possess complete data for the analysis. This might be either due to human or technical error. In case of KAMU, the same Clinic Reporting Tool is used across all the clinics. The reason for the data to be missing can be interpreted as some of the clinics failed to record to these values or the loss of data due to technical issues. Whatever might be the reason for missing data, for carrying an effective Data Envelopment Analysis we need to estimate the missing data values precisely. The accuracy of the estimated data influences the efficiency scores.

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 at least one output, Fare and Grosskopf (2002). The accuracy of the results depends on the quality and quantity of the data. The difficulty involved in replacing missing data values emanates from the fact that DEA is based on a single set of data for each attribute. The accuracy of the results directly depends on the quality and quantity of the data, since the efficiency scores are more sensitive to data errors, missing values, and data quality, Kuosmanen (2009).

The following section addresses few common methods form the literature of DEA to handle missing values.

## 3.9 Literature Review of Methods to Treat 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., (2007). This particular assumption limits the applicability of the DEA methodology to real world problems which may contain missing values either due to human errors or technical problems.

Since the problem of missing data is quite emphasized in DEA analysis, there have been different approaches reported in the literature for mitigating this problem. One such approach is the exclusion of DMU's 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 the 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 accompanied by the use of weight restrictions to reduce the impact of the missing data, Kuosmanen (2009). Some other approximation techniques such as the use of average value for replacing the missing data are also reported in the literature; however replacing multiple missing values of a single input or output variable with a single static value affects the accuracy of the calculated efficiency scores.

The other approaches for using DEA with missing values suggest interval based DEA models in which an interval range is estimated for each missing value. In this case the best suitable missing value is identified within the interval range. Another approach is to predict the best possible and least possible efficiency scores, providing an efficiency score range for DMUs with missing data (Smirlis et al., 2006). Other sophisticated methods to deal with missing values are using fuzzy membership functions developed from observational data corresponding to the missing values (Kao and Liu, 2000). This concept is similar to replacing missing values by an interval but here each value possesses a membership grade. The bounds of the interval can be determined by using statistical, experimental techniques, or expert opinions.

## 3.10 Data Envelopment Analysis Software

All the Data Envelopment Analysis models performed in this thesis are based on the software named DEA-Solver supplied along with the textbook "Data Envelopment Analysis, A Comprehensive Text with Models, Applications, References and DEA-Solver Software", Second Edition by Cooper, W.W., Seiford, L.M., and Tone, K. This software uses Microsoft Excel as base platform for the operations.

# Chapter 4 - AVERAGE RATIO METHOD

This chapter introduces the Average Ratio Method (ARM) methodology to determine the missing values. This method is considered as a basic method to address the issue of missing values in DEA. It is based on the concept of correlation, and does not involve much mathematical computations. This chapter presents greater details about this methodology, step by step procedure, its advantages, and limitations. This particular methodology will be illustrated in greater detail using simple example. Effectiveness of this methodology will be evaluated by comparing the outcomes of ARM methodology with other basic methods in literature to replace the missing values. The proposed methodology is used to evaluate the efficiency of 41 KAMU clinics with sparse data, and identifies the level of utilization of the resources for providing better services.

This chapter is structure as follows. Section 4.1, introduces the Average Ratio Methodology and the concepts behind it. Section 4.2, presents the step by step procedure of Average Ratio Methodology and the reason behind the name. Section 4.3, illustrates the step by step procedure of Average Ratio Methodology using an example dataset. Section 4.4, evaluates the effectiveness of DEA methodology using a dataset obtained from the literature. Section 4.5, once again illustrates the ARM to estimate the missing input and output value of KAMU data. Finally the DEA results for KAMU dataset will be presented in Section 4.6 and section 4.7, provides the conclusions.

## 4.1 Introduction to Methodology (ARM)

There exist many basic methods in the literature to analyze DEA with sparse data. Replacing the missing values in a particular variable with their mean or median, are few examples. There are many other such basic methods which were discussed in the previous chapter, section 3.9. The common nature identified among the literature methods are, most of them depends on the historical data or statistical methods using the average values of a particular variable, or using the distribution of the data. The basic idea behind the development of this methodology is to use the relationship between the variables.

Correlation is one such common statistical technique which can determine the linear relationship between two variables. The degree of relationship between the two variables is directly proportional to the value of correlation. Consider two variables **X** and **Y**, then the correlation between them can be calculated as follows.

$$Correlation\ = \frac{N\sum XY-\ (\sum X)(\sum Y)}{\sqrt{[N\sum X^2-(\sum X)^2][N\sum Y^2-(\sum Y)^2]}}$$

$$where\ N = \#\ of\ observations$$

The correlation value will always end up between -1.0 and +1.0. Correlation value of negative 1 implies the perfect negative correlation while a value of positive 1 implies the perfect positive correlation, and a value of zero implies the lack of correlation. If the correlation value is negative then, if one variable increases then other variable decreases and vice versa. If the correlation is positive then, if one variable increases then other variable also increases and vice versa.

Consider an example dataset with height and weight of students being used for a particular data analysis. The relation between the two variables can be computed using the correlation function. If any particular data value, within these two variables, is missing then the based on the relation between them other can be estimated. Consider one more example dataset with Medical Staff Expenses, an important variable for the data analysis, which possess missing values. In order to estimate the missing values we need to identify a variable which possess good relation with Medical Staff Expenses. The amount of expenditure spent on the Medical Staff definitely has a direct relationship with number of medical staff employees (doctors and nurses). If a particular value is missing within the variable Medical Staff Expenses then based on the relationship, with Number of Medical Staff, it can be estimated. These examples provide insight that relationship between two variables can be used to estimate the missing values.

In order to estimate the missing values based on the concept of correlation there are a list of few requirements that need to possess:

**Additional variables**: other than the list of inputs and outputs being used for analysis we need data with additional variables to determine the missing values. In order to determine these missing values precisely we need additional variables to be highly correlated to the variables with missing values. These additional variables with high correlation will not be a part of the analysis. They only support the process to determine the missing values. Correlation coefficient is used to estimate the relationship between the variables. Greater the value of correlation, greater the accuracy of results. The additional variable need to possess complete data or at least data corresponding to the missing data values in missing variable should be available.

## 4.2 Average Ratio Method (ARM)

This particular section illustrates the step by step procedure of Average Ratio Method to determine the missing values. ARM is based on the concept of correlation between two variables, with one being input or output variable with missing values and the second being an additional variable with high correlation to the first. This methodology is named as Average Ratio Method, since the ratio values are calculated for all the corresponding pairs between these two variables. The average value of all such ratio is the key factor to determine the missing values.

Consider a dataset of $n$ DMUs with $m$ inputs and $s$ outputs. Let the input and output data variables for $DMU_j$ be $(I_{j1}, I_{j2}, \dots, I_{jm})$ and $(O_{j1}, O_{j2}, \dots, O_{js})$ respectively with missing values. Consider any input variable $(I_{ji})$ $(i = 1,2, \dots, m)$ or output variable $(O_{jr})$ $(r = 1,2, \dots, s)$ which posses single or multiple missing values. Step by step by procedure of Average Ratio Method (ARM) to estimate the missing values is stated below.

1. Pick any one of the input or output variable which possess single or multiple missing values, let's say X$_1$.
2. Identify any other input or output or additional variable which satisfy the following characteristics, let's say X$_2$.
    - It should possess complete data or at least no missing values corresponding to the DMUs with already existing missing values in X$_1$.
    - There should be a high value of correlation between these two variables.

3. Considering the two attributes $X_1$ and $X_2$: calculate the ratio for all the corresponding pairs between these two variables. Such a ratio can be calculated either as $(X_1/X_2)$ or $(X_2/X_1)$.
4. The average value of all such corresponding pairs is called as Average Ratio Value $( ARV = \sum_{i=1}^{n}(X_{i1}/X_{i2}) )$.
5. If the average ratio value is obtained by $(X_1/X_2)$, then missing values in the attribute of $X_1$ is replaced by multiplying the corresponding value of the variable $X_2$ with the Average Ratio Value $(X_2 * ARV)$.
6. If the average ratio value is obtained by $(X_2/X_1)$, then missing values in the attribute of $X_1$ is replaced by dividing the corresponding value of the variable $X_2$ with the Average Ratio Value$(X_2/ARV)$.
7. We replace all the missing values in each of the input or output attributes by repeating the above process till all the missing values are replaced.

**Advantages:**

The Average Ratio Method possess following advantages:

- Less computational intensive
- Basic formulations of the DEA is not affected
- Need not run the model multiple times, unlike the fuzzy and interval models
- Determines unique values, multiple missing values are not replaced by a single average value
- Ability to determine extreme missing values, unlike methods which focus on averages
- Estimates crisp value for replacement of missing values, unlike an interval range

**Limitations:**

This methodology cannot be applied effectively if any additional variables with good correlation do not exist.

## 4.3 Numerical Example

This particular section explains the step by step procedure of ARM presented in the previous section more effectively using an example dataset. The example dataset considered possess few missing values to illustrate the ARM procedure.

Consider a complete dataset of 10 DMUs consuming a single input $I_1$ to produce a single output, $O_1^*$. $O_1^*$ represents the output variable with missing values. The output values of DMUs E and H, represented as $E_{O1}$ and $H_{O1}$ respectively, are assigned as missing values. Consider an additional variable $A_1$ which possess good relationship with the output variable $O_1^*$ and high value of correlation. The Table 4.1 presents the example dataset with the additional variable.

**Table 4-1: Numerical Example**

| DMU | $I_1$ | $O^*_1$ | | $A_1$ |
|---|---|---|---|---|
| A | 35 | 63 | | 61 |
| B | 45 | 67 | | 72 |
| C | 29 | 44 | | 68 |
| D | 26 | 86 | | 89 |
| E | 23 | ? | | 50 |
| F | 20 | 36 | | 51 |
| G | 26 | 50 | | 62 |
| H | 40 | ? | | 62 |
| I | 41 | 99 | | 88 |
| J | 10 | 58 | | 73 |

The correlation value between additional variable $A_1$ and output variable $O_1^*$ for the above dataset is 0.89. The ratio for all the pairs between the additional variable $A_1$ and output variable $O_1^*$ is calculated. The average value of all such ratios (ARV) is also calculated. The Table 4.2 presents the calculation procedures.

**Table 4-2: Calculation of Average Ratio Value and Missing Values**

| DMU | $O^*_1$ | $A_1$ | $O^*_1/A_1$ | Missing Values |
|---|---|---|---|---|
| A | 63 | 61 | 1.033 | |
| B | 67 | 72 | 0.931 | |
| C | 44 | 68 | 0.647 | |
| D | 86 | 89 | 0.966 | |

| | | | | |
|---|---|---|---|---|
| E | ? | 50 | 0 | $E_{O1}$ = 0.876 x 50 = 44 |
| F | 36 | 51 | 0.706 | |
| G | 50 | 62 | 0.806 | |
| H | ? | 62 | 0 | $H_{O1}$ = 0.876 x 62 = 54 |
| I | 99 | 88 | 1.125 | |
| J | 58 | 73 | 0.795 | |
| | **ARV = 0.876** | | | |

Since the ratio values are calculated as $O_1^*/A_1$ the missing values $E_{O1}$ and $H_{O1}$ are calculated as Average Ratio Value times the corresponding value of additional variable, $A_1$.

## 4.4 Effectiveness of Average Ratio Method

This particular section evaluates the effectiveness of the Average Ratio Method using one of the datasets from the literature studies. Table 4.3 represents the dataset of 5 DMUs consuming single input, X to produce two outputs $Y_1$, $Y_2$, considered from the literature, Kusomanen (2009). The numerical values, in Table 4.3, within brackets () represent the values assumed to be missing. These values will be determined based on the Average Ratio Method and the results are compared against the outcomes of the methodology adopted by Kusomanen (2009).

**Table 4-3: Dataset from the literature**

| **DMU** | **X** | **$Y_1$** | **$Y_2$** |
|---|---|---|---|
| A | 1 | 15 | 45 |
| B | 1 | (20) | 60 |
| C | 1 | 35 | 40 |
| D | 1 | (45) | 30 |
| E | 1 | 50 | 10 |

Kusomanen (2009) presents different methods to evaluate the data with missing values along with his research work. He identifies that replacing the missing input values by a large value and replacing the missing output values by zero will provide better efficiency score. He compared the obtained efficiency scores with other methods such as removing the DMUs or input and output variables with missing values. He found that efficiency scores obtained by his method are closer to the efficiency scores that would be obtained with actual data.

**Table 4-4: Comparison of the Efficiency Scores for Different Approaches**

| | **Efficiency Indices** | | | | |
|---|---|---|---|---|---|
| **DMU** | **(I) $T_{IDEAL}$** | **(II) $T_{DMU}$** | **(III) $T_{XY}$** | **(IV) $T_{UB}$** | **(V) $T_{ARM}$** |
| A | 0.75 | 1 | 0.75 | 0.89 | 0.8 |
| B | 1 | -- | 1.00 | 1 | 1 |
| C | 0.98 | 1 | 0.66 | 1 | 1 |
| D | 1 | -- | 0.50 | 0.5 | 0.98 |
| E | 1 | 1 | 0.17 | 1 | 1 |
| **MAD** | | **0.09** | **0.33** | **0.132** | **0.018** |
| **MAPE** | | **11.79** | **33.13** | **14.14** | **2.14** |

$T_{IDEAL}$ represent ideal efficiency when the complete data is present, $T_{DMU}$ represent the efficiency score when DMUs with missing data is removed, $T_{XY}$ represent the efficiency score when missing input or output variables removed, and $T_{UB}$ represents the efficiency score obtained using Kusomanen (2009). The efficiency scores computed using the estimated dataset by ARM method, is represented as $T_{ARM}$. The Table 4.4 compares the calculated efficiency scores for the above dataset using the different approaches. The effectiveness of these methods is evaluated by calculating the Mean Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE) between the actual and recovered efficiency scores. Calculated values are shown in Table 4.4.

Comparison of the efficiency scores from Table 4.4 demonstrates that ARM method is the clear winner. ARM is more advantageous when compared to the other methods since the ARM efficiency scores are closer to actual efficiency scores. Whereas we are not eliminating any DMU or input and output variable due to missing values and we are providing the efficiency scores for all the observation with precisely estimated values.

## 4.5 Estimating Missing Values of KAMU dataset

This section presents the final list of selected inputs and outputs with missing values. This section once again illustrates the Average Ratio Method by step by step procedure for determining the missing values for one input and one output variable. Missing values of other inputs and outputs are estimated on the same lines. Additional variables based on which the missing values of each input and output are replaced, and the correlation between them is also presented in this chapter. The complete dataset of inputs and outputs estimated based on the ARM is shown at the end of this section.

Final list of inputs and outputs identified after the preparing the dataset are presented below.

**Table 4-5: Final List of Inputs and Outputs**

| Key No | Input Variables | Key No | Output Variables |
|---|---|---|---|
| I1 | Medical Staff Expenses | O1 | Medicaid Charged |
| I2 | Facility Expenses | O2 | Self Pay Charged |
| I3 | Nurses F.T.E | O3 | Total Users |
| I4 | Administration F.T.E | | |

The list of inputs and outputs with missing values is presented in Table 4.6

**Table 4-6: List of Inputs and Outputs with Missing Values**

| Key No | I1 | I2 | I3 | I4 | O1 | O2 | O3 |
|---|---|---|---|---|---|---|---|
| 1 | 243466 | 44434 | 1 | 2 | 141318 | 180907 | 4303 |
| 2 | 153403 | 72221 | 1 | 1 | 27547 | 44553 | 2517 |
| 3 | 659744 | 70778 | 7.1 | 6.9 | 2424900 | 206302 | 6241 |
| 4 | 923910 | 1290800 | 5.56 | 11.1 | 2320627 | 2112791 | 12327 |
| 5 | 102752.8 | 28316.88 | 2.5 | 1 | 16620.01 | 17340 | 3982 |
| 6 | | 81100 | | | 277243 | 218994 | 2170 |
| 7 | 439839 | 191703 | 1 | 3 | | 77408 | 1206 |
| 8 | 17500 | 29000 | | 2 | | 15400 | 403 |
| 9 | 5255 | | | | | | 193 |
| 10 | | | | 1.5 | | | 666 |
| 11 | 1758133 | 465252 | 15.05 | 8.31 | 331613 | 1517362 | 8836 |
| 12 | 116546 | 189773 | 2 | 3.5 | 100700 | 372400 | 3220 |
| 13 | 941641 | 839882 | 2.41 | 8.91 | 257966 | 1832648 | 12532 |
| 14 | 219960 | 62000 | 3 | 2 | | | 1565 |
| 15 | 134771 | 130192 | 2.2 | 3 | | 37580 | 1589 |
| 16 | 265938 | 238043 | 2.4 | 4.43 | 68813 | 328460 | 1967 |
| 17 | 151438 | 210638 | 2 | 1.5 | | | 1954 |
| 20 | 1912377 | 1181758 | 10.78 | 14.49 | 1323691 | 3553475 | 20811 |
| 21 | 151961 | 141439 | | 1.86 | | 664125 | 3746 |
| 22 | 471560 | 409398 | 4 | 0.75 | 137545 | 457770 | 4893 |
| 23 | 504433 | 118085 | 3.6 | 3.7 | | 44857 | 1616 |
| 24 | | 72751 | | | | 766705 | 3398 |
| 25 | 97518 | | 1 | 0.5 | 50670 | 40837 | 1705 |
| 26 | 190148 | 48000 | | | 64131 | 339696 | 8012 |
| 28 | 137520 | | 2 | 1.5 | 42937 | 63021 | 2731 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 29 | 513998 | 744343 | 3.03 | 10.7 | 153073 | 559788 | 5074 |
| 31 | | | 2 | | 7343 | 450734 | 3192 |
| 32 | 172726 | 96835 | | 3 | | 33903 | 646 |
| 33 | 785527 | 889349 | 5.43 | 5.9 | 1585107 | 985674 | 6068 |
| 34 | 1305490 | 1119579 | 12.98 | 10.9 | 768776 | 1330411 | 6738 |
| 35 | 346756 | 104722 | 1 | 0.8 | 169647 | 180967 | 1262 |
| 37 | 349051.6 | 111495.8 | 3 | 2 | 53404.83 | 58152.03 | 6063 |
| 38 | 238915 | 158717 | 1 | 1.5 | | | 3193 |
| 39 | 781415 | 508496 | 2 | 6 | 188309 | 1656490 | 6173 |
| 40 | 861516 | 86678 | 2.5 | 3 | 70051 | 109446 | 3005 |
| 41 | 498324 | 389730 | 1 | | 72429 | 88174 | 1533 |
| 42 | 462062 | 426698 | 2 | 4.75 | 149784 | 588643 | 3564 |

### *4.5.1 Replacing Missing Input Values of Medical Staff Expenses*

This particular section elaborates the step by step procedure of ARM for replacing the missing values of input variable “Medical Staff Expenses”.

**Step 1:** Identify all the additional variables that have the good correlation value with the input variable “Medical Staff Expenses”.

Self Pay Charged, Nurses FTE, Administration FTE, and Total Visits are the variables that have good correlation values of 0.825, 0.845, 0.823, and 0.852 respectively with “Medical Staff Expenses”. Except the variable “Total Visits” all other variables are part of the analysis, hence the additional variable “Total Visits” is selected for the replacing the missing values.

**Step 2:** Find ratios for all the pairs between the variables “Medical Staff Expenses” and “Total Visits” and average value for all the ratios. The results are shown in Table 4.7.

**Step 3:** The missing values of the input variable “Medical Staff Expenses” are replaced by multiplying the Average Ratio Value with the corresponding values of the additional variable “Total Visits”. The recovered values are shown in Table 4.7.

**Table 4-7: Replacing Missing Values in Medical Staff Expenses**

| Key No | I1 | Total Visits | I1/Total Visits | Recovered | Complete Data |
|---|---|---|---|---|---|
| 1 | 243466 | 13308 | 18.295 | | 243466.0 |
| 2 | 153403 | 5366 | 28.588 | | 153403.0 |
| 3 | 659744 | 16110 | 40.952 | | 659744.0 |

| | | | | | |
|---|---|---|---|---|---|
| 4 | 923910 | 38857 | 23.777 | | 923910.0 |
| 5 | 102752.8 | 5973 | 17.203 | | 102752.8 |
| 6 | | 5677 | | = 39.271 x 5677 | **222940.3** |
| 7 | 439839 | 11404 | 38.569 | | 439839.0 |
| 8 | 17500 | 606 | 28.878 | | 17500.0 |
| 9 | 5255 | 420 | 12.512 | | 5255.0 |
| 10 | | 1377 | | = 39.271 x 1377 | **54075.9** |
| 11 | 1758133 | 26012 | 67.589 | | 1758133.0 |
| 12 | 116546 | 7080 | 16.461 | | 116546.0 |
| 13 | 941641 | 32111 | 29.325 | | 941641.0 |
| 14 | 219960 | 7785 | 28.254 | | 219960.0 |
| 15 | 134771 | 3758 | 35.862 | | 134771.0 |
| 16 | 265938 | 5313 | 50.054 | | 265938.0 |
| 17 | 151438 | 4936 | 30.680 | | 151438.0 |
| 20 | 1912377 | 55556 | 34.423 | | 1912377.0 |
| 21 | 151961 | 8586 | 17.699 | | 151961.0 |
| 22 | 471560 | 9085 | 51.905 | | 471560.0 |
| 23 | 504433 | 8930 | 56.487 | | 504433.0 |
| 24 | | 8822 | | =39.271 x 8822 | **346446.9** |
| 25 | 97518 | 4479 | 21.772 | | 97518.0 |
| 26 | 190148 | 8012 | 23.733 | | 190148.0 |
| 28 | 137520 | 5395 | 25.490 | | 137520.0 |
| 29 | 513998 | 12278 | 41.863 | | 513998.0 |
| 31 | | 8863 | | =39.271 x 8863 | **348057.0** |
| 32 | 172726 | 4657 | 37.090 | | 172726.0 |
| 33 | 785527 | 31487 | 24.948 | | 785527.0 |
| 34 | 1305490 | 19281 | 67.709 | | 1305490.0 |
| 35 | 346756 | 2824 | 122.789 | | 346756.0 |
| 37 | 349051.6 | 16887 | 20.670 | | 349051.6 |
| 38 | 238915 | 6717 | 35.569 | | 238915.0 |
| 39 | 781415 | 25833 | 30.249 | | 781415.0 |
| 40 | 861516 | 18876 | 45.641 | | 861516.0 |
| 41 | 498324 | 3897 | 127.874 | | 498324.0 |
| 42 | 462062 | 10739 | 43.027 | | 462062.0 |
| **Average Ratio Value** | | | **39.271** | | |

### *4.5.2 Replacing the Missing Output Values of Self Pay Charged*

This particular section elaborates the step by step procedure of ARM for replacing the missing values of output variable “Self Pay Charged”.

**Step 1:** Identify all the additional variables that have the good correlation value with the variable "Self Pay Charged".

Total Users, Total Visits are the additional variables that have good correlation value of 0.906, 0.908 respectively with "Self Pay Charged". Since "Total Users" is part of the analysis the additional variable "Total Visits" is selected for replacing the missing values.

**Step 2:** Find ratios for all the pairs between the variables "Self Pay Charged" and "Total Visits" and average value for all the ratios. The results are shown in Table 4.8.

**Step 3:** The missing values of the output variable "Self Pay Charged" are replaced by multiplying the Average Ratio Value with the corresponding values of the additional variable "Total Visits". The recovered values are shown in Table 4.8.

**Table 4-8: Replacing Missing Values in Self Pay Charged**

| Key No | O2 | Total Visits | O2/ Total Visits | Recovered | Complete Dataset |
|---|---|---|---|---|---|
| 1 | 180907 | 13308 | 13.594 | | 180907 |
| 2 | 44553 | 5366 | 8.303 | | 44553 |
| 3 | 206302 | 16110 | 12.806 | | 206302 |
| 4 | 2112791 | 38857 | 54.373 | | 2112791 |
| 5 | 17340 | 5973 | 2.903 | | 17340 |
| 6 | 218994 | 5677 | 38.576 | | 218994 |
| 7 | 77408 | 11404 | 6.788 | | 77408 |
| 8 | 15400 | 606 | 25.413 | | 15400 |
| 9 | | 420 | | = 36.51 x 420 | **15334.32** |
| 10 | | 1377 | | = 36.51 x 1377 | **50274.67** |
| 11 | 1517362 | 26012 | 58.333 | | 1517362 |
| 12 | 372400 | 7080 | 52.599 | | 372400 |
| 13 | 1832648 | 32111 | 57.072 | | 1832648 |
| 14 | | 7785 | | = 36.51 x 7785 | **284232.61** |
| 15 | 37580 | 3758 | 10.000 | | 37580 |
| 16 | 328460 | 5313 | 61.822 | | 328460 |
| 17 | | 4936 | | = 36.51 x 4936 | **180214.79** |
| 20 | 3553475 | 55556 | 63.962 | | 3553475 |
| 21 | 664125 | 8586 | 77.350 | | 664125 |
| 22 | 457770 | 9085 | 50.387 | | 457770 |
| 23 | 44857 | 8930 | 5.023 | | 44857 |
| 24 | 766705 | 8822 | 86.908 | | 766705 |
| 25 | 40837 | 4479 | 9.117 | | 40837 |

| | | | | | |
|---|---|---|---|---|---|
| 26 | 339696 | 8012 | 42.398 | | 339696 |
| 28 | 63021 | 5395 | 11.681 | | 63021 |
| 29 | 559788 | 12278 | 45.593 | | 559788 |
| 31 | 450734 | 8863 | 50.856 | | 450734 |
| 32 | 33903 | 4657 | 7.280 | | 33903 |
| 33 | 985674 | 31487 | 31.304 | | 985674 |
| 34 | 1330411 | 19281 | 69.001 | | 1330411 |
| 35 | 180967 | 2824 | 64.082 | | 180967 |
| 37 | 58152.03 | 16887 | 3.444 | | 58152.03 |
| 38 | | 6717 | | = 36.51 x 6717 | **245239.62** |
| 39 | 1656490 | 25833 | 64.123 | | 1656490 |
| 40 | 109446 | 18876 | 5.798 | | 109446 |
| 41 | 88174 | 3897 | 22.626 | | 88174 |
| 42 | 588643 | 10739 | 54.814 | | 588643 |
| **Average Ratio Value** | | | **36.510** | | |

Missing values in other input and output variables are estimated based on the same lines. The list of additional variables and their correlation values are shown in the Table 4.9.

**Table 4-9: Corresponding Additional Variables for replacing Missing Values**

| Key No | Variable with Missing Values | Additional Variable | Correlation |
|---|---|---|---|
| I1 | Medical Staff Expenses | Total Visits | 0.852 |
| I2 | Facility Expenses | Total Visits | 0.833 |
| I3 | Nurses F.T.E | Medical Staff Expenses | 0.846 |
| I4 | Administration F.T.E | Administration Expenses<br>Total Visits | 0.870<br>0.818 |
| O1 | Medicaid Charged | Total Visits | 0.610 |
| O2 | Self Pay Charged | Total Visits | 0.908 |
| O3 | Total Users | No Missing Values | |

The complete dataset of inputs and outputs obtained using the Average Ratio Method is shown in Table 4.10. The values in bold represent the data replaced using Average Ratio Method.

**Table 4-10: List of Input and Output Variables with Complete Data**

| Key No | I1 | I2 | I3 | I4 | O1 | O2 | O3 |
|---|---|---|---|---|---|---|---|
| 1 | 243466.0 | 44434.0 | 1 | 2.00 | 141318.00 | 180907.00 | 4303 |
| 2 | 153403.0 | 72221.0 | 1 | 1.00 | 27547.00 | 44553.00 | 2517 |
| 3 | 659744.0 | 70778.0 | 7.1 | 6.90 | 2424900.00 | 206302.00 | 6241 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 4 | 923910.0 | 1290800.0 | 5.56 | 11.10 | 2320627.00 | 2112791.00 | 12327 |
| 5 | 102752.8 | 28316.9 | 2.5 | 1.00 | 16620.01 | 17340.00 | 3982 |
| 6 | **222940.3** | 81100.0 | **1.8** | **1.81** | 277243.00 | 218994.00 | 2170 |
| 7 | 439839.0 | 191703.0 | 1 | 3.00 | **274650.90** | 77408.00 | 1206 |
| 8 | 17500.0 | 29000.0 | **1** | 2.00 | **14594.74** | 15400.00 | 403 |
| 9 | 5255.0 | **11134.8** | **1** | **0.50** | **10115.17** | **15334.32** | 193 |
| 10 | **54075.9** | **36506.1** | **1** | 1.50 | **33163.30** | **50274.67** | 666 |
| 11 | 1758133.0 | 465252.0 | 15.05 | 8.31 | 331613.00 | 1517362.00 | 8836 |
| 12 | 116546.0 | 189773.0 | 2 | 3.50 | 100700.00 | 372400.00 | 3220 |
| 13 | 941641.0 | 839882.0 | 2.41 | 8.91 | 257966.00 | 1832648.00 | 12532 |
| 14 | 219960.0 | 62000.0 | 3 | 2.00 | **187491.87** | **284232.61** | 1565 |
| 15 | 134771.0 | 130192.0 | 2.2 | 3.00 | **90506.67** | 37580.00 | 1589 |
| 16 | 265938.0 | 238043.0 | 2.4 | 4.43 | 68813.00 | 328460.00 | 1967 |
| 17 | 151438.0 | 210638.0 | 2 | 1.50 | **118877.31** | **180214.79** | 1954 |
| 20 | 1912377.0 | 1181758.0 | 10.78 | 14.49 | 1323691.00 | 3553475.00 | 20811 |
| 21 | 151961.0 | 141439.0 | **1.2** | 1.86 | **206782.94** | 664125.00 | 3746 |
| 22 | 471560.0 | 409398.0 | 4 | 0.75 | 137545.00 | 457770.00 | 4893 |
| 23 | 504433.0 | 118085.0 | 3.6 | 3.70 | **215067.75** | 44857.00 | 1616 |
| 24 | **346446.9** | 72751.0 | **2.8** | **1.60** | **212466.70** | 766705.00 | 3398 |
| 25 | 97518.0 | **118744.3** | 1 | 0.50 | 50670.00 | 40837.00 | 1705 |
| 26 | 190148.0 | 48000.0 | **1.5** | **1.00** | 64131.00 | 339696.00 | 8012 |
| 28 | 137520.0 | **143028.7** | 2 | 1.50 | 42937.00 | 63021.00 | 2731 |
| 29 | 513998.0 | 744343.0 | 3.03 | 10.70 | 153073.00 | 559788.00 | 5074 |
| 31 | **348057.0** | **234970.0** | 2 | **4.00** | 7343.00 | 450734.00 | 3192 |
| 32 | 172726.0 | 96835.0 | **1.4** | 3.00 | **112157.95** | 33903.00 | 646 |
| 33 | 785527.0 | 889349.0 | 5.43 | 5.90 | 1585107.00 | 985674.00 | 6068 |
| 34 | 1305490.0 | 1119579.0 | 12.98 | 10.90 | 768776.00 | 1330411.00 | 6738 |
| 35 | 346756.0 | 104722.0 | 1 | 0.80 | 169647.00 | 180967.00 | 1262 |
| 37 | 349051.6 | 111495.8 | 3 | 2.00 | 53404.83 | 58152.03 | 6063 |
| 38 | 238915.0 | 158717.0 | 1 | 1.50 | **161770.44** | **245239.62** | 3193 |
| 39 | 781415.0 | 508496.0 | 2 | 6.00 | 188309.00 | 1656490.00 | 6173 |
| 40 | 861516.0 | 86678.0 | 2.5 | 3.00 | 70051.00 | 109446.00 | 3005 |
| 41 | 498324.0 | 389730.0 | 1 | **8.70** | 72429.00 | 88174.00 | 1533 |
| 42 | 462062.0 | 426698.0 | 2 | 4.75 | 149784.00 | 588643.00 | 3564 |

In the previous section we discussed the importance of the normalization of the data prior to Data Envelopment Analysis. The normalized data values of the complete data are presented in Table 4.11. The values in bold represent the data replaced using Average Ratio Method.

**Table 4-11: Normalized Inputs and Outputs**

| Key No | I1 | I2 | I3 | I4 | O1 | O2 | O3 |
|---|---|---|---|---|---|---|---|
| 1 | 0.127 | 0.034 | 0.066 | 0.138 | 0.058 | 0.051 | 0.207 |
| 2 | 0.080 | 0.056 | 0.066 | 0.069 | 0.011 | 0.013 | 0.121 |
| 3 | 0.345 | 0.055 | 0.472 | 0.476 | 1.000 | 0.058 | 0.300 |
| 4 | 0.483 | 1.000 | 0.369 | 0.766 | 0.957 | 0.595 | 0.592 |
| 5 | 0.054 | 0.022 | 0.166 | 0.069 | 0.007 | 0.005 | 0.191 |
| 6 | **0.117** | 0.063 | **0.120** | **0.125** | 0.114 | 0.062 | 0.104 |
| 7 | 0.230 | 0.149 | 0.066 | 0.207 | **0.113** | 0.022 | 0.058 |
| 8 | 0.009 | 0.022 | **0.066** | 0.138 | **0.006** | 0.004 | 0.019 |
| 9 | 0.003 | **0.009** | **0.066** | **0.035** | **0.004** | **0.004** | 0.009 |
| 10 | **0.028** | **0.028** | **0.066** | 0.104 | **0.014** | **0.014** | 0.032 |
| 11 | 0.919 | 0.360 | 1.000 | 0.573 | 0.137 | 0.427 | 0.425 |
| 12 | 0.061 | 0.147 | 0.133 | 0.242 | 0.042 | 0.105 | 0.155 |
| 13 | 0.492 | 0.651 | 0.160 | 0.615 | 0.106 | 0.516 | 0.602 |
| 14 | 0.115 | 0.048 | 0.199 | 0.138 | **0.077** | **0.080** | 0.075 |
| 15 | 0.070 | 0.101 | 0.146 | 0.207 | **0.037** | 0.011 | 0.076 |
| 16 | 0.139 | 0.184 | 0.159 | 0.306 | 0.028 | 0.092 | 0.095 |
| 17 | 0.079 | 0.163 | 0.133 | 0.104 | **0.049** | **0.051** | 0.094 |
| 20 | 1.000 | 0.916 | 0.716 | 1.000 | 0.546 | 1.000 | 1.000 |
| 21 | 0.079 | 0.110 | **0.080** | 0.128 | **0.085** | 0.187 | 0.180 |
| 22 | 0.247 | 0.317 | 0.266 | 0.052 | 0.057 | 0.129 | 0.235 |
| 23 | 0.264 | 0.091 | 0.239 | 0.255 | **0.089** | 0.013 | 0.078 |
| 24 | **0.181** | 0.056 | **0.186** | **0.110** | **0.088** | 0.216 | 0.163 |
| 25 | 0.051 | **0.092** | 0.066 | 0.035 | 0.021 | 0.011 | 0.082 |
| 26 | 0.099 | 0.037 | **0.100** | **0.069** | 0.026 | 0.096 | 0.385 |
| 28 | 0.072 | **0.111** | 0.133 | 0.104 | 0.018 | 0.018 | 0.131 |
| 29 | 0.269 | 0.577 | 0.201 | 0.738 | 0.063 | 0.158 | 0.244 |
| 31 | **0.182** | **0.182** | 0.133 | **0.276** | 0.003 | 0.127 | 0.153 |
| 32 | 0.090 | 0.075 | **0.093** | 0.207 | **0.046** | 0.010 | 0.031 |
| 33 | 0.411 | 0.689 | 0.361 | 0.407 | 0.654 | 0.277 | 0.292 |
| 34 | 0.683 | 0.867 | 0.862 | 0.752 | 0.317 | 0.374 | 0.324 |
| 35 | 0.181 | 0.081 | 0.066 | 0.055 | 0.070 | 0.051 | 0.061 |
| 37 | 0.183 | 0.086 | 0.199 | 0.138 | 0.022 | 0.016 | 0.291 |
| 38 | 0.125 | 0.123 | 0.066 | 0.104 | **0.067** | **0.069** | 0.153 |
| 39 | 0.409 | 0.394 | 0.133 | 0.414 | 0.078 | 0.466 | 0.297 |
| 40 | 0.450 | 0.067 | 0.166 | 0.207 | 0.029 | 0.031 | 0.144 |
| 41 | 0.261 | 0.302 | 0.066 | **0.600** | 0.030 | 0.025 | 0.074 |
| 42 | 0.242 | 0.331 | 0.133 | 0.328 | 0.062 | 0.166 | 0.171 |

## 4.6 DEA Results

This section presents the application of DEA methodologies discussed so far to the complete dataset obtained using Average Ratio Method. DEA methodology can be classified into two categories, Input model and Output model. The input oriented models aims at minimizing the input consumed by the DMUs for the same target of output levels, while the output oriented models maximize the outputs produced by the DMUs for the given amount of inputs consumed. As we discussed that input oriented model is more appropriate for analysis of KAMU dataset. The results are determined for input oriented model using CCR and BCC models, and are also compared against each other, shown in Table 4.12.

**Table 4-12: Efficiency Scores**

| | **CCR Input Model** | | **BCC Input Model** | |
|---|---|---|---|---|
| **Key No** | **DEA Score** | **CCR Rank** | **DEA Score** | **BCC Rank** |
| 1 | 1.000 | 1 | 1.000 | 1 |
| 2 | 0.479 | 24 | 1.000 | 1 |
| 3 | 1.000 | 1 | 1.000 | 1 |
| 4 | 1.000 | 1 | 1.000 | 1 |
| 5 | 0.920 | 12 | 1.000 | 1 |
| 6 | 0.641 | 19 | 0.829 | 24 |
| 7 | 0.681 | 17 | 1.000 | 1 |
| 8 | 0.648 | 18 | 1.000 | 1 |
| 9 | 1.000 | 1 | 1.000 | 1 |
| 10 | 0.417 | 27 | 1.000 | 1 |
| 11 | 0.394 | 32 | 0.830 | 23 |
| 12 | 0.906 | 14 | 0.908 | 22 |
| 13 | 1.000 | 1 | 1.000 | 1 |
| 14 | 0.620 | 20 | 0.648 | 26 |
| 15 | 0.401 | 29 | 0.491 | 31 |
| 16 | 0.291 | 34 | 0.450 | 34 |
| 17 | 0.485 | 23 | 0.624 | 27 |
| 20 | 0.714 | 16 | 1.000 | 1 |
| 21 | 1.000 | 1 | 1.000 | 1 |
| 22 | 1.000 | 1 | 1.000 | 1 |
| 23 | 0.217 | 37 | 0.370 | 37 |
| 24 | 1.000 | 1 | 1.000 | 1 |
| 25 | 0.609 | 21 | 1.000 | 1 |
| 26 | 1.000 | 1 | 1.000 | 1 |

|  |  |  |  |  |
|---|---|---|---|---|
| 28 | 0.500 | 22 | 0.571 | 29 |
| 29 | 0.399 | 30 | 0.420 | 35 |
| 31 | 0.437 | 26 | 0.552 | 30 |
| 32 | 0.251 | 36 | 0.751 | 25 |
| 33 | 1.000 | 1 | 1.000 | 1 |
| 34 | 0.398 | 31 | 0.460 | 32 |
| 35 | 0.913 | 13 | 1.000 | 1 |
| 37 | 0.415 | 28 | 0.452 | 33 |
| 38 | 0.792 | 15 | 1.000 | 1 |
| 39 | 1.000 | 1 | 1.000 | 1 |
| 40 | 0.264 | 35 | 0.419 | 36 |
| 41 | 0.358 | 33 | 1.000 | 1 |
| 42 | 0.470 | 25 | 0.582 | 28 |

The DEA score of 1 implies that the particular clinic (DMU) is efficient and such group of units forms the efficient frontier for estimating the relative efficiency of other DMUs. The DEA score for other DMUs are determined with respect to the frontier formed by efficient units. Lower score indicates their larger distance from the efficient frontier and need for improvement. The inefficiency of the DMU implies that there are few production units among the peers who could produce the same amount of outputs with a lesser consumption of inputs.

Comparison between CCR and BCC efficiency scores reveals that CCR efficiency score are subset of BCC efficiency scores. If the DMU is CCR efficient then it is BCC efficient also while the converse is not true.

## 4.7 Conclusions

The Kansas Association for the Medically Underserved (KAMU) provided us with the data to evaluate the performance of 41 clinics. Four of those clinics did not have sufficient data, which are even beyond the scope for estimation, for the DEA analysis. Thus only 37 clinics participated in the analysis. Seven most appropriate input and output variables are identified to execute the DEA methodology. These procedures are accomplished successfully by the end of previous chapter.

The chapter performed the following steps to evaluate the performance of 41 clinics.

1. Replacing missing data values
2. Performing the analysis using a variety of DEA models

The primary issue associated with the data provided by KAMU is that that a large amount of data was sporadically missing, where each clinic collected a different subset of the data. This chapter proposes a new approach known as Average Ratio Method to determine the missing values based on the concept of correlation. Later the step by step procedure of this methodology is illustrated using an example dataset. The effectiveness of this methodology is evaluated by considering a dataset from the literature. The outcomes of the Average Ratio Method are compared against the outcomes of the literature. The comparison indicates ARM as the benchmark. Then the Average Ratio Methodology is used to determine the missing values of the KAMU data. DEA methodology is carried out based on the complete dataset achieved using Average Ratio Methodology.

Among the DEA models, CCR (Input oriented formulation), BCC (Input oriented formulation) are used. The variety of models helps to analyze the consistency of the results as well as ranking of the top clinics. As the result the eleven most efficient clinics using both the CCR and BCC models are shown in Table 4.13.

**Table 4-13: Efficient Clinics**

| Rank | Clinic |
|---|---|
| 1 | 39 |
| 1 | 1 |
| 1 | 33 |
| 1 | 3 |
| 1 | 4 |
| 1 | 26 |
| 1 | 24 |
| 1 | 22 |
| 1 | 21 |
| 1 | 9 |
| 1 | 13 |

The DEA methodology is more accurate in generating the benchmarks that each clinic needs to achieve, rather than the absolute ranking. We hope that the results presented in this report can be used to improve the operational aspects of KAMU clinics.

Overall this chapter presents a comprehensive study of KAMU clinics using the DEA methodology.

# Chapter 5 - DATA CLUSTERING

This chapter presents a methodology based on fuzzy clustering concepts to execute Data Envelopment Analysis with sparse input and output data. This chapter thus provides an introduction to data clustering, then to fuzzy clustering concepts. The approach presented in this chapter is based on a modified fuzzy c-means clustering using Optimal Completion Strategy (OCS) algorithm. This particular existing algorithm is sensitive to the initial values chosen to substitute missing values and also to the selected number of clusters. Therefore, this chapter 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 KAMU clinics, assuming varying levels of missing data. Values are also assumed as missing based on three common types of missing values. Results show the effect of the clustering based approach on the data recovered considering the amount and type of missing data. Moreover, this chapter shows the effect that the recovered data has on the DEA scores.

This chapter is structured as follows. Section 5.1, provides introduction to data clustering, terminology, and the basic types of clustering. Section 5.2, introduces the fuzzy c-means clustering which provides the background to the methods that will be discussed in this chapter. Section 5.3 presents literature works in the field of clustering to handle missing values. Section 5.4, illustrates the issues associated with the Optimal Completion Strategy (OCS) algorithm using an example data. Section 5.5 presents the improved version of OCS algorithm and its application to real and complete dataset of 22 KAMU clinics. Section 5.6 illustrates the effect of data recovered using clustering on DEA outcomes. Finally Section 5.7 provides the conclusions.

## 5.1 Introduction to Data Clustering

Clustering is a process of classifying data items into specific groups or clusters based on the degree of similarity between the data items. Similarity measure and coefficients play an important role in cluster analysis, since they quantify the similarity or dissimilarity between any two data items. Clustering also holds the assumption for availability of complete numerical data. Dealing with missing values in clustering is discussed in section 5.3. More details regarding the clustering methodology, models, and applications can be found in Gan et al., (2007). Cluster analysis has been applied to many fields such as health care systems Congdon (1997), Chacon and Luci (2003) and marketing Ray et al., (2005) are among many others. This section also presents the required terminology that will be used throughout this chapter.

***Notations:***

$i$ = $1,2,3,\ldots,n$, where $n$ represents the total number of observations

$j$ = $1,2,3,\ldots,d$, each observation possesses multiple attributes (d)

$u_{ik}$ = Represents membership grade of $i^{th}$ observation in $k^{th}$ cluster

$v$ = Represents cluster centers of the *c* cluster (c x *d* matrix), where $v_k$ represent cluster k

$c$ = Denotes total number clusters where, $k = 1,2,3,\ldots c$

$r$ = Represent step value or iteration number in the clustering process

$X$ = $[x_1, x_2, \ldots, x_n]^T$, Represent a data set of $n$ observations

$x_i$ = $i^{th}$ observation with *d*- dimensional data vector, for $1 \leq i \leq n$

$x_{ij}$ = $j^{th}$ attribute of $i^{th}$ observation, for $1 \leq i \leq n,\ 1 \leq j \leq d$

$X_p$ = Represent the set of $x_{ij}$values which are present in X

$X_M$ = Represent the set of $x_{ij}$values which are missing in X

$X_{Obs}$ = Represent set of entities (observations) with completely observed data (all *d* attributes)

$D_{ik}$ = Distance from $i^{th}$ observation to $k^{th}$ cluster

The interpretation of the similarity between the data items generally depends on the distance between them. Some of the common distance measures are Euclidean Distance, Manhattan Distance, Maximum Distance, Minkowski Distance, Mahalanobis Distance, and Average Distance. Most of these distance functions can be derived from Minkowski Distance, which can be stated as follows to obtain the distance between two observations **X** and **Y**.

$$d(\mathbf{x},\mathbf{y}) = \left(\sum_{j=1}^{d} |x_j - y_j|^r\right)^{1/r}, \qquad r \geq 1 \tag{5.1}$$

The Euclidean distance, Manhattan distance, and maximum distance are three specific cases of the Minkowski distance, where the Manhattan distance is defined by r = 1, Euclidean distance by r = 2, and Maximum distance is calculated using r = ∞.

Clustering algorithms can be broadly classified into hard clustering (crisp) and fuzzy clustering. Hard clustering assumes that each observation belongs to only one particular cluster group. Fuzzy clustering allows each observation to belong to more than one cluster with a certain membership value. The following Table 5.1 represents the conditions for hard clustering and fuzzy clustering, Gan et al., (2007).

**Table 5-1: Conditions for Hard and Fuzzy clustering**

| Hard Clustering (Crisp) | Fuzzy Clustering |
|---|---|
| $\mathrm{u_{ij}} \in \{0,1\}, \quad 1 \le \mathrm{i} \le \mathrm{n}, \quad 1 \le \mathrm{j} \le \mathrm{d}$<br>$\sum_{j=1}^{d} u_{ij} = 1, \quad 1 \le \mathrm{i} \le \mathrm{n}$<br>(5.2)<br>$\sum_{i=1}^{n} u_{ij} > 0, \quad 1 \le \mathrm{j} \le \mathrm{d}$ | $\mathrm{u_{ij}} \in [0, 1], \quad 1 \le \mathrm{i} \le \mathrm{n}, \quad 1 \le \mathrm{j} \le \mathrm{d}$<br>$\sum_{j=1}^{d} u_{ij} = 1, \quad 1 \le \mathrm{i} \le \mathrm{n}$<br>(5.3)<br>$\sum_{i=1}^{n} u_{ij} > 0, \quad 1 \le \mathrm{j} \le \mathrm{d}$ |

Hard clustering algorithms can be further classified into Partitional and Hierarchical clustering algorithms, with Hierarchical approaches consisting of Divisive and Agglomerative approaches.

### *5.1.1 Hierarchical Clustering Algorithms*

Hierarchical clustering algorithms are the most commonly used and can be divided into agglomerative and divisive approaches. Agglomerative clustering is a bottom up approach that starts with every single object in its own single cluster, and then repeatedly merges the closest pair of clusters according to some similarity criteria until all of the data points join a single cluster. Divisive clustering or top-down approach starts with all the objects in one cluster and repeatedly splits large clusters into smaller ones.

Agglomerative hierarchical methods include The Single Link method (Florek et al., 1951), Complete Link method (Johnson, 1967), Ward's method (Ward Jr., 1963), Group Average, Weighted Group Average, Centroid and Median methods (Jain and Dubes, 1988). Divisive methods can be sub divided into two type, monothetic and polythetic which divides the data sets into groups based on single and multiple attributes respectively. The DIANA method presented in Kaufman and Rousseeuw (1990), DISMEA (Spath, 1980), and the Edwards and Cavalli-Sforza method (1965) are a few examples of divisive hierarchical clustering algorithms.

The disadvantages of both approaches are: (a) data points that have been incorrectly grouped at an early stage cannot be reallocated and (b) different similarity measures may lead to different results.

### *5.1.2 Partitional Clustering Algorithms*

Unlike the hierarchical clustering algorithms, partitional algorithms aim at classifying the clusters at once and are based on a criterion function. The algorithm proceeds by trying to optimize the criterion function which is generally a measure of dissimilarity and thus tries to assign the cluster groups. K-means clustering by MacQueen (1967) is a common example of partitional clustering algorithms, with a fixed number of clusters known a priori. The advantage of this methodology is its ease of implementation and efficiency, while a disadvantage is the difficulty in determining the number of clusters a priori.

## 5.2 Fuzzy C Means Clustering

Fuzzy c-means (FCM) is a method of clustering which allows each entity to belong to two or more clusters. This method (developed by Dunn in 1973 and improved by Bezdek in 1981) is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$Min_{(U,v)} \left\{ J_m(U,v) = \sum_{i=1}^{n} \sum_{k=1}^{c} (\mathrm{u}_{ik})^{\mathrm{m}} \|\mathrm{x}_i - v_k\|^2 \right\}, \qquad 1 < \mathrm{m} < \infty \tag{5.4}$$

The FCM allows each entity represented by an attribute vector to belong to every cluster with a fuzzy truth value (between 0 and 1). Following are the steps of the Fuzzy C-Mean Clustering algorithm (Bezdek, 1981):

**Step 1:** Fix $\mathrm{c}\ (2 \le \mathrm{c} < n)$ and select a value for $\mathrm{m}(1 < m < \infty)$. Initialize $\mathrm{U}^{(r)}$ such that condition (5.5) is satisfied. Each step in the algorithm will be labeled as $r$ where r = 0, 1, 2……..

$$\textstyle\sum_{k=1}^{c} \mathrm{u}_{ik} = 1 \ \ \forall\, i; \ \ \sum_{i=1}^{n} \mathrm{u}_{ik} > 0 \ \ \forall\, k \tag{5.5}$$

**Step 2:** Calculate $c$ fuzzy cluster centers $\mathrm{v_k}^{\mathrm{r}}$ for each step using $\mathrm{U}^{(r)}$ and (5.6)

$$v_k = \frac{\sum_{i=1}^{n}(u_{ik})^m x_i}{\sum_{i=1}^{n}(u_{ik})^m} \quad \forall\, k = 1,..,c \tag{5.6}$$

**Step 3:** Update the initial membership function from $U^{(r)}$ to $U^{(r+1)}$ using $v_k{}^r$ and (5.7)

$$u_{ij} = \frac{1}{\sum_{K=1}^{c}\left(\frac{\|x_i - c_j\|}{\|x_i - c_k\|}\right)^{\frac{2}{m-1}}} \tag{5.7}$$

**Step 4:** If the difference between the updated and original membership matrix i.e., $\left\|U^{(r+1)} - U^{(r)}\right\| < \varepsilon_r$ then STOP; otherwise set $r = r + 1$ and return to step 2.

Note that the FCM algorithm has been somewhat generalized; and some algorithms initialize $v^{(0)}$ and check for $\left\|v^{(r+1)} - v^{(r)}\right\| < \varepsilon_r$.

## 5.3 Clustering with Missing Data

Generally methods dealing with missing data can be classified into two major approaches (Fujikawa and Ho, 2002):

(a) Pre-replacing methods, which replace missing values before the data analysis process.

(b) Embedded methods, which deal with missing values during the data analysis process.

Some of the common methods for pre- replacing missing values stated by Fujikawa and Ho, (2002) are statistics-based methods including linear regression, replacement under same standard deviation and mean-mode method. Machine learning-based methods include nearest neighbor estimator, auto associative neural network, and decision tree imputation also fall into this category. Embedded methods include case-wise deletion, lazy decision tree, dynamic path generation and some popular methods such as C4.5 and CART.

Few common clustering methods, based on the fuzzy c-means algorithm developed by Hathaway and Bezdek (2001) which are used to determine missing values are discussed below.

### *5.3.1 Whole Data Strategy (WDS)*

This particular approach is simple and valid for data sets with small proportion of missing values. Data vectors with missing values are deleted and then fuzzy c-means clustering is applied. This algorithm provides better results if less than 25% of the data points are missing. Thus, the WDS provides membership values for vectors of complete dataset only. Membership of missing data vectors need to be estimated based on nearest-prototype classification scheme using partial distances, which is presented in the following section. This method holds all the convergence properties of fuzzy c-mean clustering (Hathaway and Bezdek, 2001).

### *5.3.2 Partial Distance Strategy (PDS)*

This approach is more applicable for cases with large data sets. It is based on scaling the calculated partial distance by the quantity of data items used. Thus it reduces the influence of incomplete data values on the distance calculated.

Using this approach, the partial distance (squared Euclidean) is calculated using all available values and then scaled by reciprocal of the proportion of components used. The general formula for partial distance $D_{ik}$ is given by:

$$D_{ik} = \frac{d}{I_i}\sum_{j=1}^{d}\left(x_{ij} - v_{kj}\right)^2 I_{ij}\,, \qquad where$$

(5.8)

$$I_{ij} = \begin{cases} 0 \;\; if\; x_{ij} \epsilon\, X_M \\ 1\;\; if\; x_{ij} \epsilon\, X_p \end{cases} \; for\; 1 \leq j \leq d \;and\; 1 \leq i \leq n, and\; I_i = \sum_{j=1}^{d} I_{ij} \;\; where$$

$$X_p = \left\{x_{ij}\; for\; 1 \leq i \leq n \;and\; 1 \leq j \leq d \mid \; the\; value\; for\; x_{ij} is\; present\; in\; X\right\}$$

$$X_M = \left\{x_{ij} = unknown \;\; for\; 1 \leq i \leq n \;and\; 1 \leq j \leq d \mid \; the\; value\; for\; x_{ij} is\; missing\; in\; X\right\}$$

The partial distance strategy algorithm is obtained by making two important modifications to the FCM algorithm. (1) Calculate $D_{ik}$ for incomplete data according to equation (5.8) and (2) replace the calculation for new cluster centers by the old centers multiplied by $I_{ij}$

where $I_{ij}$ is zero for corresponding missing values. Here $v_{kj}$ represents that $j^{th}$ attribute value of the center of cluster *k*.

$$v_{kj}{}^{(r+1)} = \frac{\left(\sum_{i=1}^{n}\left(U_{ik}{}^{(r+1)}\right)^{m} I_{ij}x_{ij}\right)}{\left(\sum_{i=1}^{n}\left(U_{ik}{}^{(r+1)}\right)^{m} I_{ij}\right)}$$

(5.9)

This algorithm also holds all convergence properties of fuzzy c-mean clustering (Hathaway and Bezdek, 2001).

### *5.3.3 Optimal Completion Strategy Algorithm (OCS)*

OCS algorithm is an extension to fuzzy c-means (FCM) algorithm with an additional step to optimize the missing values over each iteration. OCS modification of FCM is referred as OCSFCM, posses all the convergence properties of FCM. At the beginning of the algorithm missing values in the dataset are replaced by some initial values. The effect of choosing different types of values can influence the results, which will be discussed in section 5.4. Missing values are considered additional variables which are estimated by minimizing the objective function of FCM. At each iteration missing values are estimated using the step 5 of the OCS algorithm. Estimated missing values are placed into the dataset at each iteration, and the algorithm continues until the termination condition of FCM (step 4) is satisfied. This algorithm is referred to as a tri-level alternating optimization, and for convergence properties refer (Hathaway et al., 2001).

The first four steps of OCS algorithm are the same as the FCM clustering algorithm, the additional step of OCS algorithm is as follows:

**Step 5:**Calculate missing values for the iteration *r+1* using equation (5.10). Place the calculated missing values into the dataset and proceed to the next iteration until the condition in step 4 (of FCM) is satisfied.

$$x_{ij}^{(r+1)} = \left[\sum_{k=1}^{c}\left(U_{ik}^{(r+1)}\right)^{m} v_{kj}^{(r+1)}\right] \Big/ \left[\sum_{k=1}^{c}\left(U_{ik}^{(r+1)}\right)^{m}\right] \ \forall\, x_{ij} \in x_{M} \qquad (5.10)$$

### *5.3.4 Nearest Prototype Strategy (NPS)*

This algorithm is a simple modification to the OCS algorithm. Here, the missing values of incomplete data item are substituted by the corresponding values of the cluster center to which the data point has highest membership degree (Hathaway and Bezdek, 2001).

In the NPS approach the additional step (Step 5) of OCS algorithm which estimates the missing values is replaced by the equation (5.11). Theoretical convergence properties of this method have not yet been proved.

$$\mathrm{x}_{\mathrm{ij}}^{(\mathrm{r}+1)} = v_{\mathrm{kj}}^{(\mathrm{r}+1)} \quad where\ D_{ik} = \min\{D_{i1}, D_{i2}, \dots\dots.., D_{ic}\}\ \forall\ \mathrm{x}_{\mathrm{ij}} \in \mathrm{x}_{\mathrm{M}} \tag{5.11}$$

## 5.4 Effect of Initial Values and Cluster Size on OCS

The previous section discussed some important algorithms to handle missing values in clustering. OCS algorithm seems to produce a better set of results since the convergence properties of this algorithm are proven. The two issues associated with optimal completion strategy (OCS) algorithm are initializing the missing values and determination of cluster size. Missing values at the beginning of the OCS algorithm need to be replaced by some initial values. This section illustrates the effect that selecting such initial values to replace the missing values will have on the final results, using an example. Consider a small dataset with 10 objects and 2 attributes taken from a real dataset, as shown in Table 5.2. Two values (10%) of the dataset are randomly assigned as missing values. Assume that $X_{21}$ and $X_{72}$ values as missing. The effect of the cluster size on the data recovered is also demonstrated using the same example.

Assumed missing values are replaced by initial values based on three different methods:

- Type 1: Missing values in each attribute are initially replaced by average value of the attribute.
- Type 2: Missing values in the dataset are initially replaced by using Average Ratio Method, discussed in the previous chapter.
- Type 3: Missing values in the dataset are initially replaced by zero.

**Table 5-2: Initial Dataset**

| | Y1 | Y2 |
|---|---|---|
| X1 | 0.127 | 0.102 |

| | | |
|---|---|---|
| X2 | **0.080** | 0.098 |
| X3 | 0.345 | 0.297 |
| X4 | 0.483 | 0.461 |
| X5 | 0.054 | 0.018 |
| X6 | 0.041 | 0.135 |
| X7 | 0.230 | **0.195** |
| X8 | 0.009 | 0.019 |
| X9 | 0.003 | 0.002 |
| X10 | 0.065 | 0.017 |

Table 5.3 presents the values placed into the dataset initially for estimating the missing values $X_{21}$, $X_{72}$ of the original dataset.

**Table 5-3: Initial Values Generated by the Three Approaches**

| Missing Values | Original Values | Type 1 | Type 2 | Type 3 |
|---|---|---|---|---|
| $X_{21}$ | 0.080 | 0.151 | 0.153 | 0.000 |
| $X_{72}$ | 0.195 | 0.128 | 0.146 | 0.000 |

Since it is difficult to determine the optimal number of clusters, we experimented with 2 to 7 clusters, considering n=8 $(1 < c < n)$ objects which possess complete data. The OCS algorithm is applied to the three different datasets, obtained by replacing the missing values, using different number of clusters. The recovered values obtained using the OCS algorithm for different number of clusters is compared to the original values using the Mean Absolute Percent Error (MAPE), as shown in Table 5.4.

**Table 5-4: Values Recovered using OCS algorithm with Different Number of Clusters**

| | Missing Values | Original Values | Different number of clusters | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | 2 Clusters | 3 Clusters | 4 Clusters | 5 Clusters | 6 Clusters | 7 Clusters |
| Type 1 | $X_{21}$ | 0.080 | 0.0496 | 0.1738 | 0.0879 | 0.0882 | 0.1260 | 0.1256 |
| | $X_{72}$ | 0.195 | 0.3678 | 0.1064 | 0.2948 | 0.2255 | 0.2072 | 0.1794 |
| | | **MAPE** | *63.31* | *81.34* | *30.53* | ***12.95*** | *31.88* | *32.50* |
| Type 2 | $X_{21}$ | 0.080 | 0.0496 | 0.1738 | 0.0880 | 0.0879 | 0.1260 | 0.1201 |
| | $X_{72}$ | 0.195 | 0.3679 | 0.1064 | 0.2948 | 0.2153 | 0.2316 | 0.2028 |
| | | **MAPE** | *63.33* | *81.34* | *30.59* | ***10.14*** | *38.13* | *27.06* |
| Type 3 | $X_{21}$ | 0.080 | 0.0496 | 0.0471 | 0.0879 | 0.0733 | 0.0432 | 0.0419 |
| | $X_{72}$ | 0.195 | 0.3690 | 0.2906 | 0.2954 | 0.1214 | 0.1600 | 0.1232 |
| | | **MAPE** | *63.62* | *45.08* | *30.68* | ***23.06*** | *31.97* | *42.22* |

The results demonstrate the influence of the initial values as well as the number of clusters on the missing values generated using the OCS approach. The results show that the missing values are best estimated using the Average Ratio Method with 5 clusters (50% of the total number of data objects, n=10). Thus we suggest the use of Average Ratio Method (ARM) to estimate the initial values prior to the application of the OCS algorithm. There is no good way to determine the optimal number of clusters which can produce the best estimates of the missing values. Thus determination of the number of clusters is left to the choice of the user. Drawing inferences from the results, we suggest choosing the number of clusters as 40 to 60% of total number of objects in the dataset.

## 5.5 Using the OCS Algorithm for Data Recovery

This section presents an application of the Optimal Completion Strategy algorithm using a real and complete dataset. The data is taken from a research project which aims at determining the productivity of 41 clinics in Kansas with 225 attributes, with the intention of improving the clinic’s quality and revenue. Since most clinics did not have complete data sets we reduced the data to 22 clinics with seven attributes, consisting of four input and three output variables. Table 5.5 shows the list of these inputs and outputs.

**Table 5-5: List of Inputs and Outputs**

| **Key No** | **Input Variables** | **Key No** | **Output Variables** |
|---|---|---|---|
| $I_1$ | Medical Staff Expenses | $O_1$ | Total Medical Visits |
| $I_2$ | Facility Expenses | $O_2$ | Self Pay Collected |
| $I_3$ | Administration full time employee | $O_3$ | State PC Collected |
| $I_4$ | Nurses full time employee | | |

The normalized and complete dataset is presented in Table 5.6.

**Table 5-6: Normalized Values of the Original Data**

| | Input Attributes | | | | Output Attributes | | |
|---|---|---|---|---|---|---|---|
| Key # | $I_1$ | $I_2$ | $I_3$ | $I_4$ | $O_1$ | $O_2$ | $O_3$ |
| 1 | 0.1273 | 0.1022 | 0.1380 | 0.0665 | 0.2909 | 0.0397 | 0.1463 |
| 4 | 0.4831 | 0.4606 | 0.7661 | 0.3694 | 0.4576 | 0.2980 | 0.4504 |
| 5 | 0.0537 | 0.0177 | 0.0690 | 0.1661 | 0.1129 | 0.0075 | 0.1701 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 7 | 0.2300 | 0.1950 | 0.2070 | 0.0665 | 0.2455 | 0.0596 | 0.1874 |
| 11 | 0.9193 | 0.4436 | 0.5735 | 1.0000 | 0.4740 | 0.5013 | 0.6058 |
| 12 | 0.0609 | 0.2636 | 0.2416 | 0.1329 | 0.1548 | 0.1278 | 0.1536 |
| 13 | 0.4924 | 0.6900 | 0.6149 | 0.1601 | 0.3583 | 1.0000 | 0.3437 |
| 14 | 0.1150 | 0.5303 | 0.1380 | 0.1993 | 0.1702 | 0.0143 | 0.1170 |
| 15 | 0.0705 | 0.0117 | 0.2070 | 0.1462 | 0.0821 | 0.0396 | 0.1178 |
| 16 | 0.1391 | 0.0804 | 0.3057 | 0.1595 | 0.1145 | 0.0810 | 0.2140 |
| 17 | 0.0792 | 0.1985 | 0.1035 | 0.1329 | 0.0937 | 0.0390 | 0.2068 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 0.7163 | 1.0000 | 0.6349 | 0.3870 |
| 22 | 0.2466 | 0.2659 | 0.0518 | 0.2658 | 0.1751 | 0.1703 | 0.3189 |
| 23 | 0.2638 | 0.1861 | 0.2554 | 0.2392 | 0.1786 | 0.0325 | 0.1158 |
| 29 | 0.2688 | 0.3750 | 0.7384 | 0.2013 | 0.2684 | 0.2248 | 0.2166 |
| 33 | 0.4108 | 0.9466 | 0.4072 | 0.3608 | 0.6018 | 0.2867 | 0.1581 |
| 34 | 0.6827 | 0.5379 | 0.7522 | 0.8625 | 0.4215 | 0.3779 | 0.5858 |
| 35 | 0.1813 | 0.2148 | 0.0552 | 0.0665 | 0.0617 | 0.0174 | 0.1755 |
| 38 | 0.1249 | 0.1621 | 0.1035 | 0.0665 | 0.1500 | 0.1321 | 0.1097 |
| 39 | 0.4086 | 0.3235 | 0.4141 | 0.1329 | 0.5293 | 0.6085 | 1.0000 |
| 40 | 0.4505 | 0.1931 | 0.2070 | 0.1661 | 0.4126 | 0.3260 | 0.1755 |
| 42 | 0.2416 | 0.2875 | 0.3278 | 0.1329 | 0.1952 | 0.2388 | 0.2627 |

The effectiveness of the OCS algorithm in recovering the missing values is evaluated by assuming various levels of data missing, ranging from 10% to 40%. In addition we assumed four different patterns of missing values including:

a) Randomly missing values. These values do not follow any pattern.
b) Missing values are centered around the attribute's average.
c) The values missing consist of extreme low and extreme high values only. Thus the 10% missing values consist of 5% of the lowest values and 5% of the highest that are eliminated.
d) The values missing consist of low input values and high out values only.

Thus, a total of ten different cases are tested including 10% random, 10% average, 10% extreme, 10% low input and high output, 20% random, 20% average, 30% random, 30% average, 40% random, and 40% average values as missing.

Notation wise the randomly missing data is denoted as "Missing Completely At Random" (MCAR), the "average" values are denoted as "Missing At Random (MAR)" (since these values close to the average are still randomly selected for elimination). The values in category (c and d) are denoted as "Missing Not At Random" (MNAR)", since this selection is based on a specific criterion and is not random notation is adopted from Little and Rubin, 2002).

The 10 different cases are demonstrated using the real and complete dataset of the 22 rural clinics, where the values assumed as missing are initially replaced based on the Average Ratio Method.

The difference between the highest and lowest missing values is represented as range for each case. The range demonstrates the variability of the missing data, with a higher range implying data further away from a possible cluster center, making it harder to regenerate. The best set of recovered values for the 10 different cases is shown in Table 5.7. In this Table the recovered values are compared with the known values that were eliminated as missing. The Table also shows Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD) and the best number of clusters for each case.

**Table 5-7: Recovered Values using OCS for different cases**

| No. | Title | Range | # of Missing Values | Best # of Clusters | MAPE | MAD |
|---|---|---|---|---|---|---|
| 1 | 10% Random | 0.9603 | 16 | 12 | 52.7 | 0.1351 |
| 2 | 10% Average | 0.2622 | 16 | 11 | 50.4 | 0.1463 |
| 3 | 20% Random | 0.9463 | 32 | 15 | 55.1 | 0.1350 |
| 4 | 20% Average | 0.2945 | 32 | 11 | 45.6 | 0.1304 |
| 5 | 30% Random | 0.7517 | 47 | 18 | 68.7 | 0.1093 |
| 6 | 30% Average | 0.3333 | 47 | 11 | 44.0 | 0.1164 |
| 7 | 40% Random | 0.9883 | 62 | 14 | 89.7 | 0.1626 |
| 8 | 40% Average | 0.5339 | 62 | 11 | 48.5 | 0.1267 |
| 9 | 10% Extreme | 0.9925 | 16 | 18 | 177.3 | 0.2897 |
| 10 | 10% Low IP & High OP | 0.9883 | 16 | 18 | 186.7 | 0.2704 |

### *5.5.1 Results and Discussions*

The results in Table 5.7 show that missing values that are close to the entity's average estimated more accurately than data missing at random, or data of extreme values, especially as more data is missing.

In the case of randomly missing values the MAPE is increasing as the percentage of missing values increases as expected as shown in Figure 5.1.

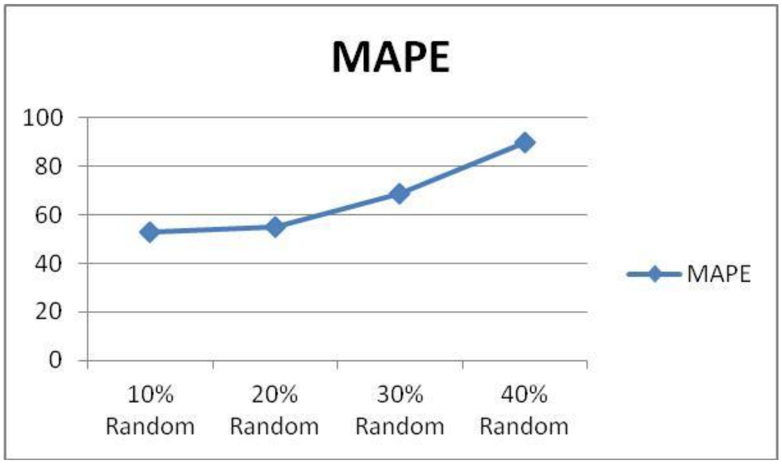

**Figure 5-1: MAPE for the case of Missing Completely At Random**

This shows that the OCS approach recovers missing values that are close to the average better than randomly missing values. The Mean Absolute Deviation of data missing at random is largely insensitive to the quantity of the missing data until the 40% mark. At that point too much data is missing which affects the accuracy of the clustering and thus data recovery as shown in Figure 5.2.

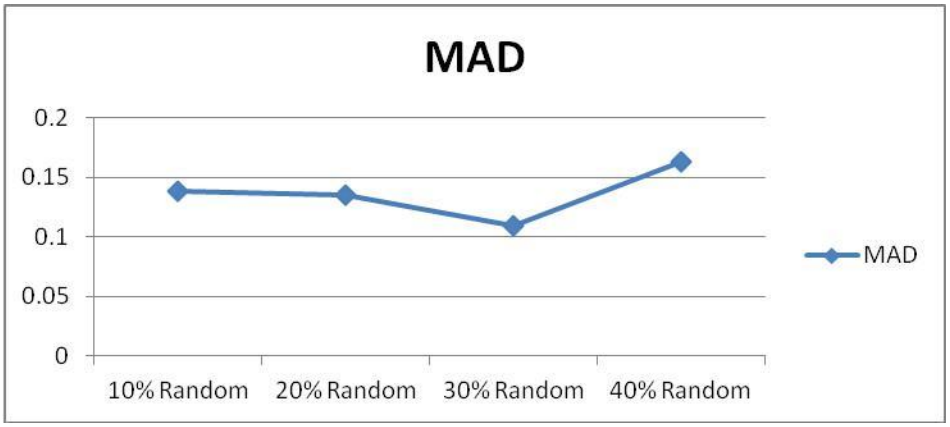

**Figure 5-2: MAD as a Function of Quantity of Missing Data**

The worst case scenarios as expected occur when the missing data is of extreme value. In this case the OCS algorithm cannot estimate the missing values accurately, since the estimates are based on the fuzzy clusters' centers.

Observation of the results from Table 5.7 also shows that under most cases the best set of missing values are recovered when the number of clusters equals about 50% of total number of observations. As the percentage of missing values increase, the preferred number of clusters increases also.

## 5.6 Data Recovery Effects on DEA Results

In the previous section we had assumed various quantities of data as missing starting from 10% to 40% under 10 different cases. (Note that the actual complete dataset of the 22 KAMU clinics with 3 inputs and 4 outputs was shown in Table 5.6) The initial set of missing values was estimated using the Average Ratio Method and the final set of missing values was generated using the OCS algorithm. Hence for the DEA analysis we have a total of 11 different datasets including 10 generated and one real and complete dataset.

The efficiency scores of the clinics based on the CCR Input oriented model are shown in Table 5.8.

The Table shows the actual efficiency of each clinic using the complete data set. Also, the Table shows the calculated efficiency using the recovered data using the 10 schemes described in section 5.5. Then the difference between the "assumed" efficiency and the "real" (with actual data) is calculated using again the Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD).

**Table 5-8: Comparison of Efficiency Scores using CCR Input Model**

| DMU Key # | Original Dataset | 10% Ran | 20% Ran | 30% Ran | 40% Ran | 10% Avg | 20% Avg | 30% Avg | 40% Avg | 10% Ext | 10% LI & HO |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | **1.000** | 1.000 | 1.000 | 1.000 | 1.000 | 0.826 | 0.810 | 0.819 | 0.965 | 1.000 | 1.000 |
| 4 | **0.558** | 0.540 | 0.634 | 0.553 | 0.496 | 0.756 | 0.693 | 0.611 | 0.641 | 0.688 | 0.813 |
| 5 | **1.000** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.836 |
| 7 | **0.860** | 0.860 | 1.000 | 0.645 | 0.919 | 0.810 | 0.785 | 0.960 | 0.966 | 1.000 | 1.000 |
| 11 | **0.611** | 0.692 | 0.682 | 0.578 | 1.000 | 0.772 | 0.739 | 0.776 | 0.661 | 0.702 | 0.743 |
| 12 | **1.000** | 1.000 | 1.000 | 0.447 | 0.967 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.897 |
| 13 | **1.000** | 1.000 | 1.000 | 1.000 | 0.524 | 1.000 | 1.000 | 1.000 | 1.000 | 0.854 | 1.000 |
| 14 | **0.655** | 0.553 | 0.787 | 0.776 | 0.647 | 0.861 | 0.916 | 1.000 | 0.937 | 0.869 | 0.713 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 15 | **1.000** | 1.000 | 1.000 | 1.000 | 0.702 | 1.000 | 1.000 | 1.000 | 1.000 | 0.923 | 0.877 |
| 16 | **0.629** | 0.686 | 0.843 | 0.836 | 1.000 | 0.886 | 0.939 | 0.979 | 1.000 | 1.000 | 1.000 |
| 17 | **0.933** | 0.908 | 1.000 | 1.000 | 0.961 | 0.756 | 0.675 | 0.717 | 0.768 | 1.000 | 1.000 |
| 20 | **0.617** | 0.769 | 0.833 | 0.597 | 1.000 | 0.760 | 0.775 | 0.706 | 0.693 | 0.726 | 0.422 |
| 22 | **1.000** | 1.000 | 1.000 | 1.000 | 0.843 | 1.000 | 1.000 | 1.000 | 1.000 | 0.785 | 1.000 |
| 23 | **0.342** | 0.341 | 0.752 | 0.475 | 0.475 | 0.667 | 0.736 | 0.888 | 0.964 | 0.345 | 0.429 |
| 29 | **0.598** | 0.565 | 0.973 | 0.709 | 0.641 | 0.594 | 0.711 | 0.878 | 1.000 | 0.660 | 0.884 |
| 33 | **0.835** | 0.455 | 0.797 | 1.000 | 0.719 | 0.955 | 0.884 | 1.000 | 0.998 | 0.840 | 0.715 |
| 34 | **0.426** | 0.501 | 0.646 | 0.510 | 0.780 | 0.546 | 0.520 | 0.549 | 0.653 | 0.580 | 0.659 |
| 35 | **1.000** | 0.521 | 0.915 | 0.801 | 0.744 | 0.949 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 38 | **0.932** | 0.848 | 0.911 | 0.832 | 1.000 | 0.996 | 1.000 | 0.875 | 0.882 | 1.000 | 1.000 |
| 39 | **1.000** | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | **1.000** | 1.000 | 1.000 | 1.000 | 0.987 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 42 | **0.610** | 0.606 | 1.000 | 0.531 | 0.865 | 0.601 | 0.622 | 0.896 | 0.986 | 0.859 | 1.000 |
| **MAPE** | | **8.87** | **20.51** | **13.14** | **23.76** | **15.95** | **17.56** | **23.63** | **25.43** | **14.13** | **20.13** |
| **MAD** | | **0.068** | **0.112** | **0.095** | **0.159** | **0.094** | **0.102** | **0.134** | **0.137** | **0.096** | **0.127** |

### *5.6.1 DEA Results and Discussions*

The results from Table 5.8 show that generally the efficiency scores deviate from the real ones as more data is missing, as shown in Figure 5.3.

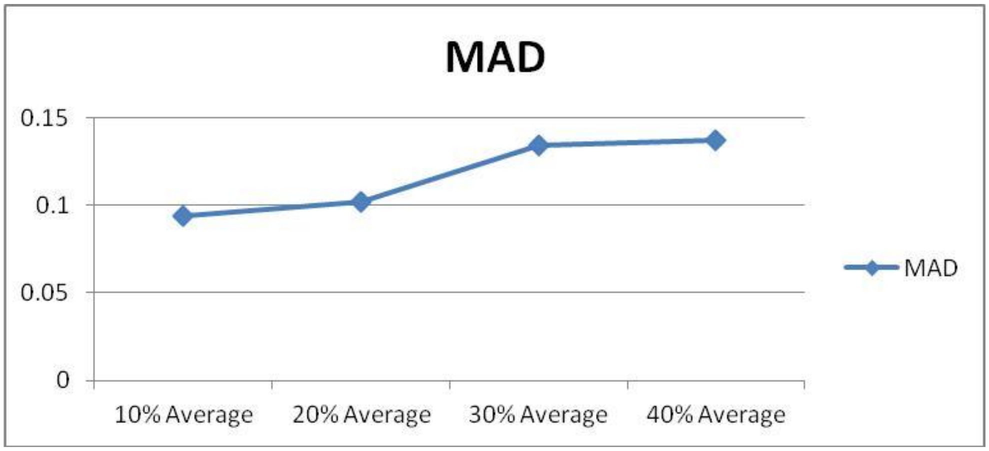

**Figure 5-3: Error in Efficiency Scores as a Function of Missing Data Quantity**

The inferences that can be clearly identified from the results are as the percentage of missing values increases the MAPE also increases.

The interesting nature of DEA scores can be observed by comparing the efficiency scores calculated with 10% extreme and 10% lowest input and highest output missing. Generally the nature of outliers present in the data can greatly affect the results, but in the case of DEA analysis the most critical observations are with lowest inputs and highest outputs. These observations denote efficient DMUs, and when these values are replaced by averages these DMU scores are degraded.

Hence when 10% of the lowest inputs and highest output values are missing, the error presented as MAPE is equivalent to the MAPE of 20% random missing values and is quite larger than any other case in the group of 10% missing values. The MAPE for the 4 different cases under the group of 10% missing values is graphically illustrated in Figure 5.4 and is compared against 20% random missing values. This shows that the influence of lowest input and highest output missing values can be greater in the case of DEA when compared to the general extreme missing values (without distinction of input or output).

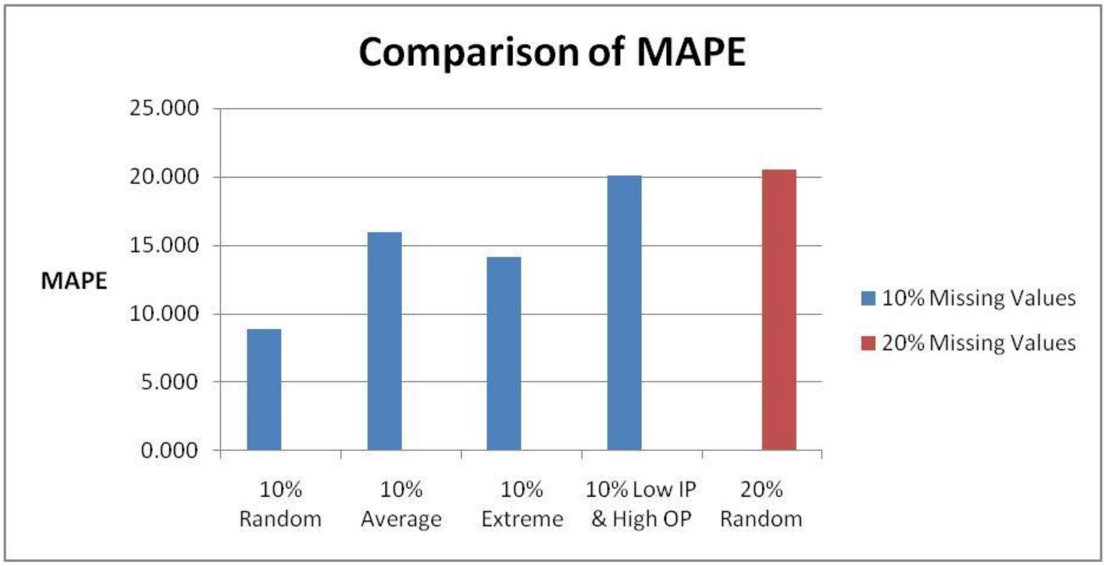

**Figure 5-4: Influence of Lowest Input & Highest Output Missing Values**

## 5.7 Conclusions

This paper provides a brief introduction to DEA Methodology, literature review of DEA in healthcare, literature review of approaches of handling missing data using DEA, and a comprehensive review of clustering approaches, and approaches of handling missing values in clustering applications.

This paper focuses on a methodology for conducting DEA analysis when some of the necessary input or output parameters are missing. The approach presented is to replace the missing values based on the data generated by a modified fuzzy c-means clustering approach enhanced by the Optimal Completion Strategy (OCS). The two major factors that could greatly affect the results are: initializing the missing values at the beginning of the clustering approach, and choosing the number of clusters. The influence of these two factors on the recovered missing values is illustrated using a short example dataset. The results suggest using the Average Ratio Method to replace the initial missing values, and to select about 50% of the total number of objects in the dataset as the number of clusters. These two recommendations are also validated using a real and complete dataset of 22 clinics.

The missing data recovery using the OCS algorithm was tested using the complete data set of the 22 clinics, with varying levels of assumed missing values, ranging from 10% to 40%. Here a total of 10 different cases were considered to test the effectiveness of the Optimal Completion Strategy (OCS) algorithm. The three basic types of missing values, Missing Completely At Random (MCAR), Missing At Random (MAR), and Missing Not At Random (MNAR) are covered under the 10 different cases. The results show that the OCS worked more effectively with values Missing At Random (MAR) (where missing values are centered around the attribute’s mean) than values Missing Completely At Random (MCAR). In the case of the MAR, the Mean Absolute Percentage Error (MAPE) is gradually decreasing as the percentage of missing values are increasing, whereas in the case of MCAR the mean absolute percentage error is gradually increasing as the percentage of missing values are increasing.

The clustering methodology generates the missing values to be used in the DEA analysis. The methodology developed here assigns the best set of recovered missing values back into the data set.

The DEA analysis performed here analyzed 22 KAMU clinics with 7 attribute, of which 3 are inputs and 4 are outputs, with varying levels of missing values. In the analysis we compared the actual efficiency scores of the clinics, calculated with the original and complete data set against the data generated using the OCS approach. The results show that the efficiency scores are fairly insensitive to the missing data – either due to a sufficiently good recovery of the data, or the averaging effect of the DEA. Even when a large amount of data is missing, the DEA results are still almost always within 0.1 of the correct efficiency score.

Thus the proposed approach is robust in the sense that the data recovered and the DEA scores generated are insensitive to the quantity of data missing!

However, when extreme data is missing, especially low input and high output values, the DEA analysis tends to underestimate the efficiencies as expected.

As a summary, this paper provides an effective and practical approach for replacing missing values needed for a DEA analysis.

# Chapter 6 - CLUSTER DISPERSION

This chapter presents a methodology to estimate missing values based on a modified fuzzy c-means clustering algorithm which takes cluster dispersion into account. This particular clustering methodology acts as an intermediate step, estimating the missing values scientifically to be used for Data Envelopment Analysis. The cluster dispersion approach reduces the likelihood of mistakenly assigning marginal data objects to larger cluster groups instead of nearer smaller clusters. The newly developed clustering approach, which takes dispersion into account, is demonstrated on a real and complete dataset of 22 KAMU clinics, assuming varying levels of missing data. Results show the effect of the clustering approach on the data recovery considering the amount of missing data. Moreover, the paper also shows the effect that the recovered data has on the DEA scores.

This chapter is structured as follows. Section 6.1, provides introduction to existing cluster dispersion methodology. Section 6.2, illustrates the issues associated with the existing cluster dispersion methodology using an example. Section 6.3 presents the new cluster dispersion methodologies. Section 6.4, presents the application of these new methodologies to real and complete dataset of 22 KAMU clinics. It also compares the effectiveness of the recovered values and convergence properties for these new methods. Section 6.5 illustrates the effect of data recovered using cluster dispersion on DEA outcomes. Finally Section 6.6 provides the conclusions.

## 6.1 Cluster Dispersion

This algorithm is an extension of Optimal Completion Strategy (OCS), discussed in the previous chapter, taking cluster dispersion into account. Himmelspach and Conrad (2010) suggest that general clustering approaches to estimate missing values work well for uniformly distributed datasets but not for real datasets that do not have uniform cluster sizes. OCS algorithm estimates missing values solely based on distances between observations and their cluster centers, hence remote data objects can be biased by cluster size.

Extending the OCS algorithm by taking cluster dispersion into account, this new algorithm by Himmelspach and Conrad (2010) is named as Fuzzy C-Means Algorithm for

Incomplete Data based on Cluster Dispersion (FCMCD). FCMCD updates the new membership function $u_{ik}{}^{*}$ taking cluster dispersion into account by computing it. This cluster dispersion, $S_k$ of a cluster $v_k$ is defined as squared average distance of data objects to their cluster centers, considering only entities with complete data, as shown in equation (6.1). ‘f’ represents the attribute values of the corresponding observation. The primary difference between calculating the FCMCD and the OCS is the usage of the cluster dispersion value, $S_k$.

$$S_k = \frac{1}{|v_k \cap X_{obs}| - 1} \sum_{x_j \in v_k \cap x_{obs}} \sum_{f \in f_{obs}} (x_i.f - v_k.f)^2 \tag{6.1}$$

Where $x_j \in v_k \, if \; and \; only \; if \; u_{kj} = \max\{u_{1j}, \dots \dots \dots \dots \dots, u_{cj}\} \; and \; |v_k \cap X_{obs}| \; \geq 2$

$X_{obs} \; represent \; set \; of \; completely \; observed \; data \; items$

$f_{obs} \; represent \; set \; of \; completely \; observed \; features$

The FCMCD algorithm, an extension of OCS and FCM, can be obtained by modifying Step 3 of Fuzzy C-Means algorithm, discussed in the previous chapter, in the following way:

**Step 3’:** This step is the primary difference between OCS and FCMCD, the process of updating the membership function where the later takes the cluster dispersion into account. Updating the membership function of the $i^{th}$ observation to cluster k, $u_{ik}{}^{*}$, using cluster dispersion is defined as.

$$u_{ik}{}^{*} = \frac{\left( S_k D_{ik}{}^{1/(1-m)} \right)}{\left( \sum_{k=1}^{c} \left( S_k \; D_{ik}{}^{\frac{1}{1-m}} \right) \right)} \tag{6.2}$$

Calculate new set of cluster centers $v$ and estimate missing values using equation 6.3. For more details of FCMCD refer to Himmelspach and Conrad, 2010. Note that convergence properties of this particular method are not discussed.

$$x_{ij}^{(r+1)} = \left[\sum_{k=1}^{c}\left(u*_{ik}^{(r+1)}\right)^{m} v_{kj}^{(r+1)}\right] \Big/ \left[\sum_{k=1}^{c}\left(u*_{ik}^{(r+1)}\right)^{m}\right] \quad \forall\, x_{ij} \in x_M \tag{6.3}$$

## 6.2 Issues associated with FCMCD

This particular section further concentrates on FCMCD and illustrates the issues associated in application of this algorithm to estimate missing values. The primary issue lies in computing the cluster dispersion value, $S_k$, which further influences the updating the membership matrix. Reviewing the dispersion, $S_k$, in equation (6.1), illustrates that it is mainly based on observations with complete data ($X_{Obs}$). There is a possibility that there are too many missing values in the dataset leaving no potential observations with complete data. However this possibility can happen even in cases with smaller percentage of missing data. If this probable case happens then it leaves no good way for computing the cluster dispersion value ($S_k$) and also updating the membership matrix.

Consider a real and complete dataset of 22 observations with 7 attributes, which will be presented later in section 6.5. Assume 10% of data is missing, which implies that potentially 16 values in the dataset can be missing. There is a possibility that 16 observations (out of 22) each can have a single missing value, which leaves 6 potential observations with complete data. Based on experimental results calculating the dispersion based on 6 observations, estimating the missing values, and trying to classify a group of 22 observations can lead to incorrect analysis. Potentially if 14% of data is assumed as missing then each observation will have a single missing value and cannot be used for computing the dispersion value. Based on the experimental results the FCMCD algorithm could fail to converge for even smaller percentage of missing values depending on the size of dataset and structure of missing values.

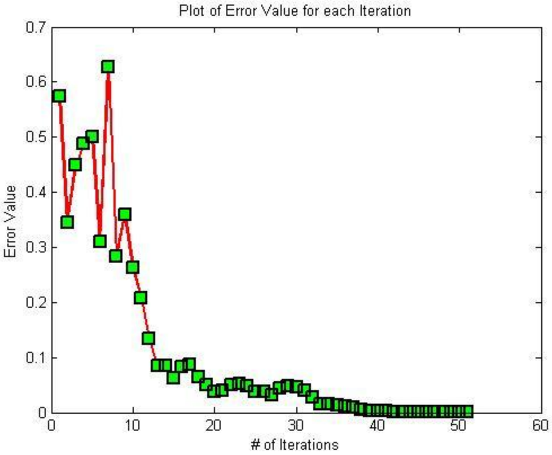

**Figure 6-1: Repetition of Error Value with 2 Clusters**

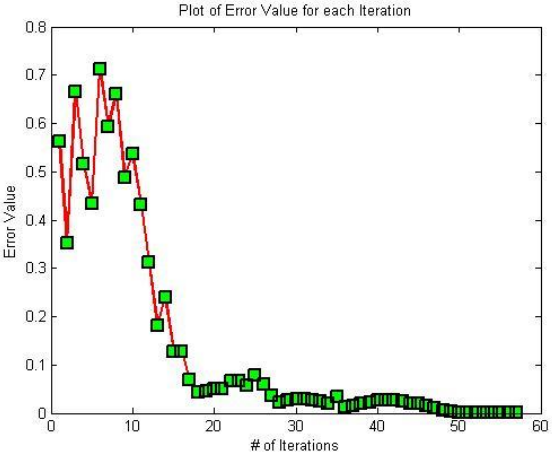

**Figure 6-2: Repetition of Error Value with 3 Clusters**

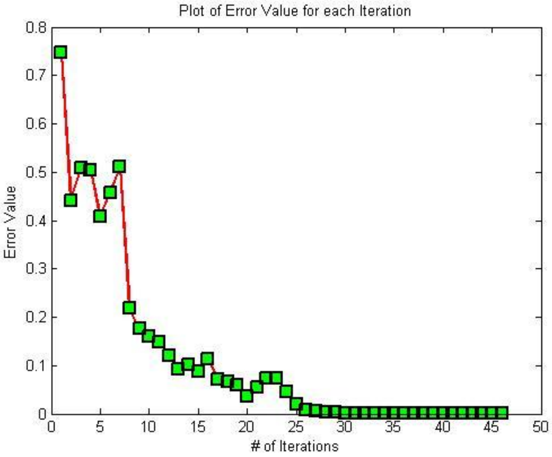

**Figure 6-3: Repetition of Error Value with 4 Clusters**

The second issue that we discovered during the experimentation process is that the FCMCD algorithm failed to converge when increasing the number of clusters. During the experiments the algorithm successfully converged to a final solution with up to 4 clusters. The iteration of the error value for cluster size from 2 to 4 is shown in Figures 6.1 to 6.3 respectively. When the number of clusters is increased to 5 or more, the algorithm failed to converge. It is evident that after a few iterations the error $\varepsilon_r$ starts to follow a pattern as shown in Figure 6.4.

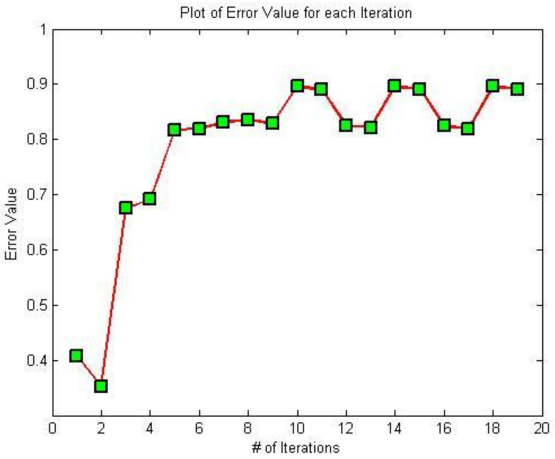

**Figure 6-4: Repetition of Error Value with 5 Clusters**

When looking at the updated membership matrix we can see that data points shift from one cluster to another repeatedly. The updated membership matrix is shown in Table 6.1 for 10$^{th}$, 11$^{th}$, and 12$^{th}$, iterations. Highlighted cells in Table 6.1 represent the maximum membership value of a particular observation in a particular cluster center. Maximum membership value of a particular cluster keeps iterating between cluster 5 and 2 in this example, showing that the algorithm does not converge to a stable fuzzy cluster.

**Table 6-1: Repetition of Membership Matrix (Cells Highlighted)**

| | 10th Iteration | | | | | 11th Iteration | | | | | 12th Iteration | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Key No** | **C1** | **C2** | **C3** | **C4** | **C5** | **C1** | **C2** | **C3** | **C4** | **C5** | **C1** | **C2** | **C3** | **C4** | **C5** |
| **1** | 0.291 | 0.028 | 0.011 | 0.027 | 0.644 | 0.027 | 0.898 | 0.012 | 0.029 | 0.033 | 0.257 | 0.027 | 0.012 | 0.029 | 0.674 |
| **4** | 0.038 | 0.001 | 0.053 | 0.886 | 0.022 | 0.004 | 0.029 | 0.058 | 0.909 | 0.001 | 0.056 | 0.001 | 0.080 | 0.831 | 0.032 |
| **5** | 0.116 | 0.013 | 0.009 | 0.018 | 0.845 | 0.027 | 0.835 | 0.024 | 0.049 | 0.065 | 0.113 | 0.013 | 0.010 | 0.021 | 0.845 |
| **7** | 0.490 | 0.033 | 0.010 | 0.028 | 0.439 | 0.030 | 0.928 | 0.007 | 0.018 | 0.016 | 0.421 | 0.038 | 0.011 | 0.032 | 0.497 |

| | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **11** | 0.020 | 0.001 | 0.917 | 0.047 | 0.015 | 0.002 | 0.017 | 0.933 | 0.048 | 0.001 | 0.024 | 0.001 | 0.903 | 0.054 | 0.018 |
| **12** | 0.471 | 0.027 | 0.009 | 0.025 | 0.468 | 0.044 | 0.895 | 0.010 | 0.027 | 0.025 | 0.419 | 0.033 | 0.009 | 0.028 | 0.512 |
| **13** | 0.189 | 0.005 | 0.196 | 0.489 | 0.121 | 0.019 | 0.165 | 0.230 | 0.581 | 0.006 | 0.194 | 0.005 | 0.195 | 0.488 | 0.119 |
| **14** | 0.513 | 0.015 | 0.034 | 0.093 | 0.346 | 0.068 | 0.716 | 0.051 | 0.142 | 0.023 | 0.467 | 0.018 | 0.035 | 0.105 | 0.376 |
| **15** | 0.120 | 0.014 | 0.008 | 0.017 | 0.841 | 0.026 | 0.843 | 0.020 | 0.043 | 0.067 | 0.117 | 0.013 | 0.008 | 0.018 | 0.844 |
| **16** | 0.268 | 0.024 | 0.011 | 0.026 | 0.671 | 0.032 | 0.872 | 0.016 | 0.037 | 0.043 | 0.245 | 0.022 | 0.011 | 0.027 | 0.695 |
| **17** | 0.088 | 0.015 | 0.004 | 0.009 | 0.884 | 0.021 | 0.871 | 0.011 | 0.025 | 0.073 | 0.084 | 0.014 | 0.004 | 0.011 | 0.887 |
| **20** | 0.091 | 0.003 | 0.619 | 0.221 | 0.067 | 0.008 | 0.080 | 0.675 | 0.234 | 0.003 | 0.090 | 0.003 | 0.640 | 0.202 | 0.066 |
| **22** | 0.532 | 0.018 | 0.018 | 0.043 | 0.389 | 0.068 | 0.813 | 0.027 | 0.065 | 0.027 | 0.453 | 0.023 | 0.020 | 0.052 | 0.452 |
| **23** | 0.487 | 0.025 | 0.015 | 0.037 | 0.437 | 0.047 | 0.874 | 0.016 | 0.040 | 0.024 | 0.422 | 0.029 | 0.016 | 0.041 | 0.492 |
| **29** | 0.327 | 0.007 | 0.086 | 0.417 | 0.164 | 0.036 | 0.282 | 0.116 | 0.556 | 0.010 | 0.328 | 0.007 | 0.086 | 0.423 | 0.156 |
| **33** | 0.232 | 0.005 | 0.131 | 0.499 | 0.133 | 0.023 | 0.194 | 0.155 | 0.621 | 0.007 | 0.219 | 0.005 | 0.128 | 0.521 | 0.127 |
| **34** | 0.038 | 0.001 | 0.817 | 0.119 | 0.026 | 0.003 | 0.032 | 0.838 | 0.125 | 0.001 | 0.040 | 0.001 | 0.814 | 0.116 | 0.029 |
| **35** | 0.220 | 0.028 | 0.009 | 0.021 | 0.722 | 0.023 | 0.902 | 0.010 | 0.026 | 0.040 | 0.207 | 0.026 | 0.009 | 0.024 | 0.735 |
| **38** | 0.166 | 0.028 | 0.005 | 0.013 | 0.789 | 0.020 | 0.906 | 0.007 | 0.017 | 0.050 | 0.142 | 0.027 | 0.005 | 0.013 | 0.813 |
| **39** | 0.252 | 0.007 | 0.182 | 0.376 | 0.183 | 0.027 | 0.261 | 0.230 | 0.473 | 0.010 | 0.243 | 0.007 | 0.188 | 0.379 | 0.182 |
| **40** | 0.656 | 0.011 | 0.027 | 0.076 | 0.230 | 0.106 | 0.674 | 0.052 | 0.148 | 0.020 | 0.626 | 0.013 | 0.027 | 0.082 | 0.252 |
| **42** | 0.887 | 0.005 | 0.006 | 0.022 | 0.081 | 0.258 | 0.619 | 0.024 | 0.083 | 0.015 | 0.962 | 0.002 | 0.002 | 0.008 | 0.027 |

For more information on cluster dispersion refer to Appendix C, which explains the cluster dispersion using an example.

## 6.3 Optimal Completion Strategy based on Cluster Dispersion

This section illustrates three new approaches for estimating the missing values based on Optimal Completion Strategy with Cluster Dispersion. The cluster dispersion values are computed using all the observations in the dataset unlike the existing method which considers only observations with complete data. These approaches try to achieve similarity among the cluster sizes reducing the opportunity for formation of few large cluster groups. Three different methods for calculating the cluster dispersion are presented below and the step by step procedure used to update the membership matrix is explained at the end.

**Method 1:**

This method tries to achieve uniformity among the clusters using Euclidean distance. The cluster dispersion value, $S_k$, of a cluster *k* is defined as the average Euclidean distance of each observation in the dataset to the corresponding cluster center.

$$S_k = \frac{1}{n}\left[\sum_{i=1}^{n} d_{euc}(\mathbf{x}_i, \mathbf{c}_k)\right] \;\forall\, k = 1, \dots, c$$

(6.4)

$$\text{Where } d_{euc}(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{d}\left(x_j - y_j\right)^2\right]^{1/2}$$

**Method 2:**

This method considers also the influence of membership grade, thus the cluster dispersion value, $S_k$, of a cluster *k* is defined as the weighted average of membership times the square of Euclidean distance

$$S_k = \left[\sum_{i=1}^{n} u_{ik} * \left(d_{euc}(\mathbf{x}_i, \mathbf{c}_k)\right)^2\right] \Big/ \left[\sum_{i=1}^{n} u_{ik}\right], \;\forall\, k = 1, \dots, c$$

(6.5)

$$\text{Where } d_{euc}(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{d}\left(x_j - y_j\right)^2\right]^{1/2}$$

**Method 3:**

Cluster dispersion plays an important role in updating the new membership matrix, having observations closer to the cluster center emphasized and ones further from a cluster center pushed to choose a nearer cluster center. Here the function that supports this influence is achieved by computing the ratio of membership to distance. Observations closer to the cluster center have shorter distance and high membership value while observations away from cluster center have larger distance and lowest membership value. The ratio of membership to distance for observations closer to cluster center will have a higher value, and for observations far away from cluster center will have a lower value. The higher value helps emphasizing closer observations and push remote observations away from the cluster towards a nearer cluster.

The cluster dispersion of cluster $(v_k, \forall\, k = 1, \dots, c)$ is defined as:

$$S_k = \frac{\sum_{i=1}^{n}\left(\frac{u_{ik}}{d_{ik}}\right)}{\sum_{i=1}^{n}(u_{ik})}$$

(6.6)

$$\text{Where } d_{ik} = \mathrm{d_{euc}}(\mathbf{x}, \mathbf{y}) = \left[\sum_{j=1}^{d}\left(\mathrm{x_j} - \mathrm{y_j}\right)^2\right]^{1/2}$$

The following is procedure of Optimal Completion Strategy algorithm based on Cluster Dispersion (OCSCD). Based on the cluster dispersion method chosen the cluster dispersion values, $S_k$, may vary however equation (6.9) and the rest of the procedure remains the same.

**Step 1:** Fix $\mathrm{c}$ $(2 \leq \mathrm{c} < n)$ and select a value for $\mathrm{m}(1 < m < \infty)$. Initialize $\mathrm{U}^{(\mathrm{r})}$ such that condition (6.7) is satisfied. Each step in the algorithm will be labeled as $r$ where r = 0, 1, 2……..

$$\sum_{k=1}^{c} u_{\mathrm{ik}} = 1 \;\; \forall\, i; \;\; \sum_{i=1}^{n} u_{\mathrm{ik}} > 0 \;\; \forall\, k \tag{6.7}$$

**Step 2:** Calculate $c$ fuzzy cluster centers $\mathrm{v_k}^{\mathrm{r}}$ for each step using $\mathrm{U}^{(\mathrm{r})}$ and (6.8)

$$v_{\mathrm{k}} = \frac{\sum_{\mathrm{i=1}}^{\mathrm{n}}(u_{\mathrm{ik}})^{\mathrm{m}}\mathrm{x_i}}{\sum_{\mathrm{i=1}}^{\mathrm{n}}(u_{\mathrm{ik}})^{\mathrm{m}}} \quad \forall\, \mathrm{k} = 1, .., \mathrm{c} \tag{6.8}$$

**Step 3:** Update the initial membership function from $\mathrm{U}^{(\mathrm{r})}$ to $\mathrm{U}^{(\mathrm{r+1})}$ using $v_{\mathrm{k}}{}^{\mathrm{r}}$ and $S_k$

$$u_{ik}{}^{*} = \frac{\left(S_k\ D_{ik}{}^{1/(1-m)}\right)}{\left(\sum_{k=1}^{c}\left(S_k\ D_{ik}{}^{\frac{1}{1-m}}\right)\right)} \tag{6.9}$$

**Step 4:**If the difference between the updated and original membership matrix i.e., $\left\|\mathrm{U}^{(\mathrm{r+1})} - \mathrm{U}^{(\mathrm{r})}\right\| < \varepsilon_r$ then STOP.

**Step 5:** Calculate missing values for the iteration '*r+1*' using equation (6.10). Place the calculated missing values into the dataset and proceed to the next iteration until the condition in step 4 is satisfied.

$$\mathrm{x_{ij}^{(r+1)}} = \left[\sum_{\mathrm{k=1}}^{\mathrm{c}}\left(u_{\mathrm{ik}}^{(\mathrm{r+1})}\right)^{\mathrm{m}} v_{\mathrm{kj}}^{(\mathrm{r+1})}\right] \Big/ \left[\sum_{\mathrm{k=1}}^{\mathrm{c}}\left(u_{\mathrm{ik}}^{(\mathrm{r+1})}\right)^{\mathrm{m}}\right] \;\; \forall\, \mathrm{x_{ij}} \in \mathrm{X_M} \tag{6.10}$$

## 6.4 Using OCSCD Algorithm for Data Recovery

This section presents an application of the Optimal Completion Strategy Cluster Dispersion (OCSCD) algorithm using a real and complete dataset, and considering the three methods that determine the cluster dispersion presented in the previous section. These approaches are tested on the complete dataset obtained from the data provided KAMU clinics. Since most clinics did not have complete data sets we reduced the dataset to 22 clinics with seven attributes, consisting of four input and three output variables. The dataset for 22 clinics is complete and does not contain any missing values. Table 6.2 shows the list of these inputs and outputs.

**Table 6-2: List of Inputs and Outputs**

| Key No | Input Variables | Key No | Output Variables |
|---|---|---|---|
| $I_1$ | Medical Staff Expenses | $O_1$ | Total Medical Visits |
| $I_2$ | Facility Expenses | $O_2$ | Self Pay Collected |
| $I_3$ | Administration Full Time Employee | $O_3$ | State PC Collected |
| $I_4$ | Nurses Full Time Employee | | |

The normalized and complete dataset is presented in Table 6.3.

**Table 6-3: Normalized Values of the Original Data**

| | Input Attributes | | | | Output Attributes | | |
|---|---|---|---|---|---|---|---|
| Key No | $I_1$ | $I_2$ | $I_3$ | $I_4$ | $O_1$ | $O_2$ | $O_3$ |
| 1 | 0.1273 | 0.1022 | 0.1380 | 0.0665 | 0.2909 | 0.0397 | 0.1463 |
| 4 | 0.4831 | 0.4606 | 0.7661 | 0.3694 | 0.4576 | 0.2980 | 0.4504 |
| 5 | 0.0537 | 0.0177 | 0.0690 | 0.1661 | 0.1129 | 0.0075 | 0.1701 |
| 7 | 0.2300 | 0.1950 | 0.2070 | 0.0665 | 0.2455 | 0.0596 | 0.1874 |
| 11 | 0.9193 | 0.4436 | 0.5735 | 1.0000 | 0.4740 | 0.5013 | 0.6058 |
| 12 | 0.0609 | 0.2636 | 0.2416 | 0.1329 | 0.1548 | 0.1278 | 0.1536 |
| 13 | 0.4924 | 0.6900 | 0.6149 | 0.1601 | 0.3583 | 1.0000 | 0.3437 |
| 14 | 0.1150 | 0.5303 | 0.1380 | 0.1993 | 0.1702 | 0.0143 | 0.1170 |
| 15 | 0.0705 | 0.0117 | 0.2070 | 0.1462 | 0.0821 | 0.0396 | 0.1178 |
| 16 | 0.1391 | 0.0804 | 0.3057 | 0.1595 | 0.1145 | 0.0810 | 0.2140 |
| 17 | 0.0792 | 0.1985 | 0.1035 | 0.1329 | 0.0937 | 0.0390 | 0.2068 |
| 20 | 1.0000 | 1.0000 | 1.0000 | 0.7163 | 1.0000 | 0.6349 | 0.3870 |
| 22 | 0.2466 | 0.2659 | 0.0518 | 0.2658 | 0.1751 | 0.1703 | 0.3189 |
| 23 | 0.2638 | 0.1861 | 0.2554 | 0.2392 | 0.1786 | 0.0325 | 0.1158 |
| 29 | 0.2688 | 0.3750 | 0.7384 | 0.2013 | 0.2684 | 0.2248 | 0.2166 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 33 | 0.4108 | 0.9466 | 0.4072 | 0.3608 | 0.6018 | 0.2867 | 0.1581 |
| 34 | 0.6827 | 0.5379 | 0.7522 | 0.8625 | 0.4215 | 0.3779 | 0.5858 |
| 35 | 0.1813 | 0.2148 | 0.0552 | 0.0665 | 0.0617 | 0.0174 | 0.1755 |
| 38 | 0.1249 | 0.1621 | 0.1035 | 0.0665 | 0.1500 | 0.1321 | 0.1097 |
| 39 | 0.4086 | 0.3235 | 0.4141 | 0.1329 | 0.5293 | 0.6085 | 1.0000 |
| 40 | 0.4505 | 0.1931 | 0.2070 | 0.1661 | 0.4126 | 0.3260 | 0.1755 |
| 42 | 0.2416 | 0.2875 | 0.3278 | 0.1329 | 0.1952 | 0.2388 | 0.2627 |

The effectiveness of OCSCD algorithm in recovering the missing values is evaluated by assuming varying levels of data as missing, ranging from 10% to 30%. The algorithm is tested using the 3 different cluster dispersion methods noted as Method 1, 2 and 3. The data assumed as missing is Missing Completely At Random, (Little and Rubin, 2002) implying that data assumed missing does not follow any pattern with respect to data present or missing.

The difference between the highest and lowest of missing values is represented as the range for each case. The range demonstrates the variability of the missing data, with a higher range implying data that is more remote from a cluster center is missing.

The Mean Absolute Deviation (MAD) between the real values (assumed as missing) and recovered values using Method 1, 2 and 3 for 10%, 20%, and 30% missing values are presented in Table 6.4.

**Table 6-4: Recovered Values using OCSCD**

| | | | | **Mean Absolute Deviation** | | |
|---|---|---|---|---|---|---|
| **No.** | **Title** | **Range** | **# of Missing Values** | **Method 1** | **Method 2** | **Method 3** |
| 1 | 10% Missing | 0.9603 | 16 | 0.1287 | 0.1161 | 0.1150 |
| 2 | 20% Missing | 0.9463 | 31 | 0.1341 | 0.1277 | 0.1153 |
| 3 | 30% Missing | 0.9925 | 47 | 0.1322 | 0.1388 | 0.1223 |

### *6.4.1 Results and Discussions*

The results in Table 6.4 show that Mean Absolute Deviation (MAD) for the recovered values keep increasing as the percentage of missing data increases. The increase of MAD between 10%, 20%, and 30% missing data varies differently for each method. Graphical illustration of MAD values for method 1, 2, and 3 is shown in Figure 6.5.

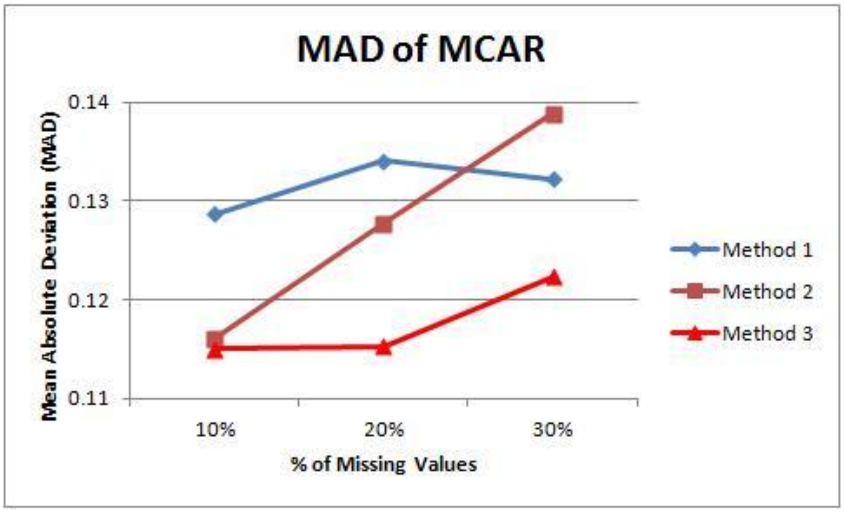

**Figure 6-5: MAD of recovered values based on Cluster Dispersion**

Based on the Mean Absolute Deviation values we can clearly infer that Method 3 is the best alternative among the 3 cluster dispersion methods to estimate the missing values. MAD value for 30% missing using Method 3 is still less when compared to 20% missing using Method 2 and 10% missing using Method 1. Method 3 is robust as the MAD does not jump much from 10% missing to 20% missing, when compared to Method 2 where the MAD keeps increasing quickly as the percentage of missing data increases.

### *6.4.2 Convergence*

Method 1 converges for any number of clusters with varying percentage of missing data as 10%, 20%, and 30%. Based on the experimental results this method is able to converge under all these cases. Method 2 converges for any number of clusters when smaller percentage of data is missing. As the percentage of missing values increase this method failed to converge especially with a higher number of clusters. The convergence for Method 3 cannot be guaranteed at higher number of clusters for any percentage of missing values. However this method is able to recover the missing values with higher fidelity with a lower number of clusters when compared to Method 1 and 2.

## 6.5 Effects of Data Recovery on DEA Results

In the previous section we assumed varying quantities of data as missing starting from 10% to 30%. As the results in section 6.4 suggest, Method 3 is the best approach towards recovering the missing data.

Here we assess the fidelity of the DEA results based on the recovered data. Thus we have a total of 3 recovered dataset with 10%, 20%, and 30% missing cases. The efficiency scores of the three recovered datasets will be compared with the efficiency scores of the real and complete dataset using DEA.

The efficiency scores of the clinics are determined based on the CCR Input oriented model. Table 6.5 shows the actual efficiency of clinics for the real and complete data set, and also the efficiency for the recovered datasets described in section 6.4. Then the difference between the "real" (with actual data) and "assumed" efficiency is calculated using the Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD).

**Table 6-5: Comparison of Efficiency Scores obtained using CCR Input Model**

| DMU Key # | Original Dataset | 10% Missing | 20% Missing | 30% Missing |
|---|---|---|---|---|
| 1 | **1.000** | 0.7059 | 1.0000 | 1.0000 |
| 4 | **0.558** | 0.6178 | 0.5656 | 0.5651 |
| 5 | **1.000** | 1.0000 | 1.0000 | 1.0000 |
| 7 | **0.860** | 0.9283 | 0.9953 | 0.9598 |
| 11 | **0.611** | 0.6225 | 0.5834 | 0.7422 |
| 12 | **1.000** | 1.0000 | 1.0000 | 1.0000 |
| 13 | **1.000** | 1.0000 | 1.0000 | 1.0000 |
| 14 | **0.655** | 0.6380 | 0.9004 | 0.8258 |
| 15 | **1.000** | 0.5794 | 0.6084 | 0.5713 |
| 16 | **0.629** | 0.6990 | 0.6912 | 0.6709 |
| 17 | **0.933** | 0.8827 | 1.0000 | 1.0000 |
| 20 | **0.617** | 0.7041 | 0.6730 | 1.0000 |
| 22 | **1.000** | 1.0000 | 1.0000 | 0.4843 |
| 23 | **0.342** | 0.5514 | 0.4706 | 0.5963 |
| 29 | **0.598** | 0.6264 | 0.7120 | 0.5506 |
| 33 | **0.835** | 0.8850 | 0.9435 | 0.5918 |
| 34 | **0.426** | 0.4056 | 0.5466 | 0.5538 |
| 35 | **1.000** | 1.0000 | 0.8619 | 0.8257 |
| 38 | **0.932** | 0.9907 | 1.0000 | 1.0000 |

| | | | | |
|---|---|---|---|---|
| 39 | **1.000** | 1.0000 | 1.0000 | 1.0000 |
| 40 | **1.000** | 1.0000 | 1.0000 | 0.7441 |
| 42 | **0.610** | 0.6049 | 0.8815 | 0.7675 |
| **MAPE** | | **9.5006** | **13.0893** | **20.3772** |
| **MAD** | | **0.0660** | **0.0883** | **0.1443** |

### *6.5.1 DEA Results and Discussions*

The results from Table 6.5 show that the efficiency scores of recovered datasets increasingly deviate from the real scores, as more data is missing. Both the Mean Absolute Percentage Error (MAPE) and Mean Absolute Deviation (MAD) keep increasing as the percentage of missing data increases. Figure 6.6 demonstrates the MAD deviation as a function of missing data quantity.

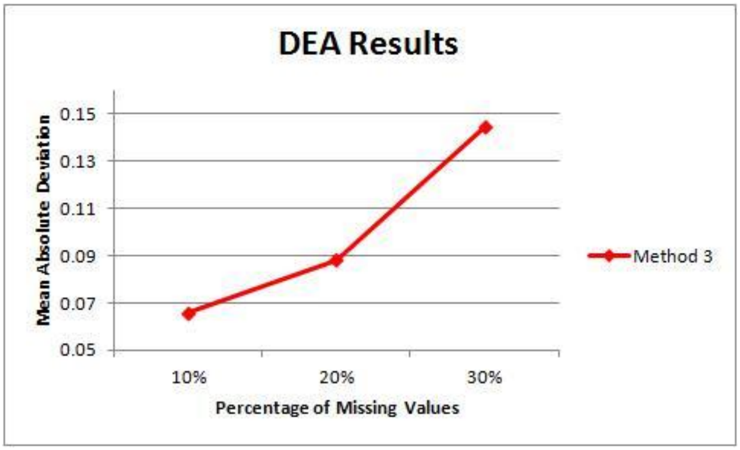

**Figure 6-6: MAD a Function of Missing Data**

The MAD of efficiency scores for 10%, 20%, and 30% missing data under Method 3 imitate the pattern observed in Figure 6.5. The difference between the 10% and 20% MAD values is minimal and increases rapidly from 20% to 30%.

## 6.6 Conclusions

This chapter focuses on an existing clustering methodology by taking cluster dispersion into account and conducting the Data Envelopment Analysis (DEA) when some critical input or output parameters are missing. The clustering approach acts as an intermediate approach to estimate the missing values for conducting DEA analysis.

The chapter highlights the importance of the cluster dispersion and its influence on estimating the missing values in the cases of real datasets. It identifies the limitations of the existing method taking cluster dispersion into account. The existing cluster dispersion method tends to fail (does not converge) even for a small percentage of missing values and with increasing number of clusters. These limitations are clearly stated and illustrated using a real dataset. Hence three new approaches to calculate the cluster dispersion are introduced.

The effectiveness of these methods is tested based on a complete data set of the 22 clinics, with varying levels of assumed missing values, ranging from 10% to 30%. The values assumed as missing are considered to be missing completely at random, and are estimated based on the three different methods proposed in this paper. Method 3 is found to be the best method among the three available methods based on the Mean Absolute Deviation between the real (assumed as missing) and recovered (estimated using the clustering).

The best set of recovered values, using Method 3, is replaced back into the dataset to perform the DEA. In the analysis we compared the actual efficiency scores of the clinics, calculated with the original and complete data set, against the data recovered using the Method 3. The results suggest that the efficiency scores are fairly insensitive to the missing data. Even when 20% data is missing, the MAD between the real and recovered efficiency scores is still less than 0.1.

The convergence properties exhibited by the three different methods under varying levels of missing data are also explored in this paper.

As a summary, this chapter provides an effective and practical approach for replacing missing values needed for a DEA analysis using Cluster Dispersion approach.

# Chapter 7 - INTERVAL DEA

This chapter presents a methodology based on interval approach to handle missing value concerns and to perform Data Envelopment Analysis (DEA). Traditionally complete data should be available to carry out DEA. In case of sparse data, generally the missing values are replaced by interval ranges estimated by experts, rest remains crisp. In most interval based DEA methods, the efficiency scores are expressed in terms of fuzzy environment but this chapter's primary focus is to express the DEA scores in terms of crisp values. Crisp efficiency scores are identified in similar lines to how DEA determines efficiency scores using the best set of weights.

The interval ranges are broken down into crisp values based on interpolations, but common alpha value is chosen for the interpolation of different ranges. The methodology in this chapter uses the concept of common alpha value for different interval ranges. The best value of alpha, for interval ranges, will be the one which endows most of the DMUs with best efficiency scores. This new interval approach is demonstrated on a real and complete dataset of 22 KAMU clinics, assuming varying levels of data as missing.

This chapter is structured as follows. Section 7.1 provides literature review to interval based DEA models to handle missing value concerns. Section 7.2, presents the formulation of the new interval approach. Section 7.3 demonstrates the application of this new interval approach using the actual clinical data for varying levels of missing values. Section 7.4 discusses the results and shows the effect of interval approach on the DEA analysis. Section 7.5 provides summary and conclusions.

## 7.1 Literature Review

Traditionally complete data should be available to carry out the Data Envelopment Analysis. However, this assumption might not be valid in all the cases. Apart from the cases with missing values there exist other types where the data is collected in form of interval data and ordinal data. DEA models developed to handle such type of data are known as Imprecise Data Envelopment Analysis (IDEA) models. The resulting DEA models turn out to be non-linear programming problems. There exists two different approaches to handle this issue; one approach is to transform the non-linear programming model to linear and to handle the interval and ordinal data. The second approach is to convert the interval or ordinal data to a set of exact data and to proceed with the standard DEA methodology. This section does not dig much into this aspect. As more of these models are applicable when all of the provided data can be classified as interval or ordinal or as ratio bounded. For more information refer to Chen and Zhu (2007), Cook and Zhu (2007). These literature works summarize most of the research work done in this area.

This section presents some of these methods which can be applied to the perspective when data possess both crisp and interval values. Kao and Liu (2000), treats the missing data with the help of interval data using fuzzy approach. Smirlis et al., (2006) treats the missing data with the help of interval data using imprecise DEA approach.

Kao and Liu (2000) adopted the concept of membership function used in fuzzy set theory, to illustrate the efficiency scores of 24 university libraries in Taiwan with 3 missing values. Triangular membership function for missing data is created using the smallest, most, and largest possible values from the observed data. Fuzzy DEA model is transformed into conventional crisp model using the concept of alpha cut. Alpha cut value indicates the corresponding input and output interval values for the membership function. Each alpha cut provides two input and two output values. In order to determine the bounds of the efficiency scores at every level of alpha, two mathematical programs are formulated to determine the minimum and maximum efficiency scores.

Maximum efficiency score occurs when lower bound of input values and upper bound of output values for the target DMU, and upper bound of input values and lower bound of output values for rest of the DMUs in the constraint set are considered. Minimum efficiency score occurs when upper bound of input values and lower bound of output values for the target DMU, and lower bound of input values and upper bound of output values for rest of the DMUs in the constraint set are considered.

For every level of alpha there exist two efficiency scores, hence increasing the alpha value in the intervals of 0.1 from zero to one we need to run the DEA methodology 22 times. The most likely efficiency score out of the 22 runs will be selected. This makes the methodology computationally intense with multiple runs. Decreasing the intervals of alpha increases the total number of runs to be performed.

The effectiveness of this methodology is not evaluated as the 3 missing values are real. Hence the most likely obtained efficiency score cannot be compared to efficiency score of real dataset. This methodology transforms interval data to crisp values and uses the standard DEA structure.

Kao and Liu (2007), presents the application of this methodology on a complete dataset assuming 1%, 2%, and 5% of the data as missing. They modified the existing triangular membership function construction process. In the new study the three vertices of the triangular membership function corresponds to the smallest, largest, and the median ranks which have appeared in other variables for the DMUs with the missing values. The new membership construction provided better results. The limitation associated with the new membership construction process is that at least one input and one output variable should possess complete data. This implies that for smaller percentage of missing values, assuming one for each variable, we need to drop few DMUs to continue DEA analysis.

Smirlis et al., (2006) adopted the concept of imprecise DEA model developed by Despotis and Smirlis (2002), to illustrate the efficiency scores of 29 secondary public schools in Greece which possess 8 missing values. The missing values are estimated within intervals based on statistical or experiential techniques. This methodology transforms the non-linear programming model to linear model, to handle the interval data. Similar to the previous model the minimum and maximum efficiency scores are determined using two mathematical programs. Based on the bounds of the interval efficiency scores, DMUs are classified into three classes. First class consists of DMUs that are efficient for any combination of inputs and outputs. Second class consists of DMUs that can act as both efficient and inefficient based on the combination of inputs and outputs. Third class consists of DMUs that are inefficient under any combination of inputs and outputs.

## 7.2 Interval Approach Formulation

This section presents new formulation based on interval approach. The missing crisp values in the dataset can be estimated with in an interval range based on expert opinion or statistical methods. The missing input and output values $x_{ji}\ and\ y_{jr}$ are expressed in terms of interval range as $[x_{ji}{}^{L},\ x_{ji}{}^{U}]$ and $[y_{jr}{}^{L},\ y_{jr}{}^{U}]$. $x_{ji}{}^{L}$ and $x_{ji}{}^{U}$ represent the lower and upper bounds of the interval range for input variables respectively. Where, $y_{jr}{}^{L}$ and $y_{jr}{}^{U}$ represent the lower and upper bounds of the interval range for output variables respectively. The interval developed for each missing value is converted into a series of crisp values using linear interpolation of the interval as follows:

$$x_{ji} = x_{ji}{}^{L} + \alpha\left(x_{ji}{}^{U} - x_{ji}{}^{L}\right) \mathit{where}\ 0 \leq \alpha \leq 1$$

$$y_{jr} = y_{jr}{}^{L} + \alpha\left(y_{jr}{}^{U} - y_{jr}{}^{L}\right) \mathit{where}\ 0 \leq \alpha \leq 1$$

The alpha value lies between zero to one. There exist several combinations to choose the alpha value for each interval range to obtain the crisp values. In this chapter we choose common alpha value for all the different interval ranges. This avoids the fuzzy environment of DEA efficiency scores. This also avoids the concept of minimum and maximum efficiency scores. Use of common alpha value reduces the overall computations by 50%. Replacing the crisp values back into the dataset, efficiency scores of the DMUs are calculated. This procedure will be repeated for different values of alpha, between zero to one.

The best value of alpha, for interval range, will be the one which endows best efficiency scores for most of the DMUs. The efficiency scores achieved for a particular value of alpha are summed. This summation value helps us in determining the best value of alpha. The alpha value corresponding to the highest summation score will be chosen as the best value of alpha and this bestows most of the DMUs with best efficiency scores. This methodology is based on the concept of transforming the interval data to precise data and making use of the standard DEA structure. The extended interval based DEA (CCR Input) model is shown below:

$$Max \quad Z = \sum_{r=1}^{s} u_r\{y_{or}{}^{L} + \alpha(y_{or}{}^{U} - y_{or}{}^{L})\}$$

$$S.t.$$

$$\sum_{i=1}^{m} v_i\{x_{0i}{}^{L} + \alpha(x_{0i}{}^{U} - x_{oi}{}^{L})\} = 1$$

$$-\sum_{i=1}^{m} v_i\{x_{ji}{}^{L} + \alpha\left(x_{ji}{}^{U} - x_{ji}{}^{L}\right)\} + \sum_{r=1}^{s} u_r\{y_{jr}{}^{L} + \alpha\left(y_{jr}{}^{U} - y_{jr}{}^{L}\right)\} \leq 0 \ \forall\, j = 1, \ldots, n$$

$$u_r, v_i \geq 0 \ \ \forall\, r = 1, \ldots, s, \ \ i = 1, \ldots, m, and\ 0 \leq \alpha \leq 1$$

We can also determine the best possible and least possible efficiency scores based on the interval range for different missing values. Clearly, for DEA analysis, a DMU will be most efficient if α→0 for input variables and α→1 for output parameters, providing a DMU with lowest possible input and highest possible output. The best efficiency score occurs when we consider the lowest input values and highest output values for the target DMU, and consider the highest input values and lowest output values for rest of the DMUs in the constraint set. The least efficiency score is possible when we consider the highest input values and lowest output values for the target DMU, and consider the lowest input values and highest output values for rest of the DMUs in the constraint set. This helps us in determining the possible interval range for DEA efficiency scores. Model 7.1 and 7.2 represent the formulation for best and least CCR input oriented models respectively, shown in Table 7.1.

**Table 7-1: Best and Least cases for CCR Input Model**

| **Best CCR Input Oriented Model** | **Least CCR Input Oriented Model** |
|---|---|
| $Max \quad Z = \sum_{r=1}^{s} u_r y_{or}^{U}$<br>$S.t.$ (7.1)<br>$\sum_{i=1}^{m} v_i x_{oi}^{L} = 1$<br>$-\sum_{i=1}^{m} v_i x_{oi}^{L} + \sum_{r=1}^{s} u_r y_{or}^{U} \le 0$<br>$-\sum_{i=1}^{m} v_i x_{ji}^{U} + \sum_{r=1}^{s} u_r y_{jr}^{L} \le 0 \;\; \forall j = 1, \ldots, n; j \neq 0$<br>$u_r, v_i \ge 0 \;\; \forall r = 1, \ldots, s, \;\; i = 1, \ldots, m$ | $Max \quad Z = \sum_{r=1}^{s} u_r y_{or}^{L}$<br>$S.t.$ (7.2)<br>$\sum_{i=1}^{m} v_i x_{oi}^{U} = 1$<br>$-\sum_{i=1}^{m} v_i x_{oi}^{U} + \sum_{r=1}^{s} u_r y_{or}^{L} \le 0$<br>$-\sum_{i=1}^{m} v_i x_{ji}^{L} + \sum_{r=1}^{s} u_r y_{jr}^{U} \le 0 \; \forall j = 1, \ldots, n; j \neq 0$<br>$u_r, v_i \ge 0 \;\; \forall r = 1, \ldots, s, \;\; i = 1, \ldots, m$ |

There exist very less probability for best and least efficiency scores cases to happen in real world and moreover these are imaginary situations. Best case model represents the upper bound for the possible efficiency scores with interval range. While the least case represents the lower bound for the possible efficiency scores with interval range. However we are not determining the multiple efficiency scores which represent the fuzzy environment. This chapter primarily focuses on determining the crisp efficiency scores.

## 7.3 Application of Interval Approach

This section presents application of the new interval approach using a real and complete dataset of 22 KAMU clinics, assuming varying levels of data as missing. Table 7.2 shows the list of these inputs and outputs. Interval ranges for the assumed missing values are constructed around the actual value based on two different approaches. In the first model intervals are constructed relatively to the actual value. In the second model intervals are fixed around the actual value. In both these cases the intervals are symmetrical around the actual value.

**Table 7-2: List of Inputs and Outputs**

| Key No | Input Variables | Key No | Output Variables |
|---|---|---|---|
| $I_1$ | Medical Staff Expenses | $O_1$ | Total Medical Visits |
| $I_2$ | Facility Expenses | $O_2$ | Self Pay Collected |
| $I_3$ | Administration Full Time Employee | $O_3$ | State PC Collected |
| $I_4$ | Nurses Full Time Employee | | |

The normalized and complete dataset is presented in Table 7.3.

**Table 7-3: Normalized Values of the Original Data**

| DMUs | Input Variables | | | | Output Variables | | |
|---|---|---|---|---|---|---|---|
| **Key No** | **$I_1$** | **$I_2$** | **$I_3$** | **$I_4$** | **$O_1$** | **$O_2$** | **$O_3$** |
| 1 | **0.1273** | 0.1022 | 0.1380 | 0.0665 | 0.2909 | ***0.0397*** | 0.1463 |
| 4 | 0.4831 | *0.4606* | **0.7661** | 0.3694 | **0.4576** | 0.2980 | 0.4504 |
| 5 | **0.0537** | 0.0177 | 0.0690 | 0.1661 | 0.1129 | 0.0075 | ***0.1701*** |
| 7 | 0.2300 | 0.1950 | **0.2070** | **0.0665** | 0.2455 | 0.0596 | **0.1874** |
| 11 | 0.9193 | **0.4436** | *0.5735* | 1.0000 | **0.4740** | 0.5013 | 0.6058 |
| 12 | **0.0609** | 0.2636 | 0.2416 | 0.1329 | ***0.1548*** | **0.1278** | 0.1536 |
| 13 | 0.4924 | **0.6900** | 0.6149 | **0.1601** | 0.3583 | 1.0000 | **0.3437** |
| 14 | *0.1150* | 0.5303 | **0.1380** | 0.1993 | 0.1702 | **0.0143** | 0.1170 |
| 15 | **0.0705** | 0.0117 | 0.2070 | 0.1462 | 0.0821 | 0.0396 | ***0.1178*** |
| 16 | 0.1391 | 0.0804 | 0.3057 | **0.1595** | **0.1145** | 0.0810 | 0.2140 |
| 17 | 0.0792 | **0.1985** | 0.1035 | 0.1329 | 0.0937 | **0.0390** | 0.2068 |
| 20 | 1.0000 | *1.0000* | 1.0000 | **0.7163** | 1.0000 | 0.6349 | **0.3870** |
| 22 | 0.2466 | **0.2659** | 0.0518 | 0.2658 | **0.1751** | 0.1703 | *0.3189* |
| 23 | **0.2638** | 0.1861 | **0.2554** | 0.2392 | 0.1786 | 0.0325 | 0.1158 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 29 | 0.2688 | 0.3750 | 0.7384 | ***0.2013*** | 0.2684 | 0.2248 | **0.2166** |
| 33 | 0.4108 | 0.9466 | **0.4072** | 0.3608 | *0.6018* | **0.2867** | 0.1581 |
| 34 | **0.6827** | 0.5379 | 0.7522 | 0.8625 | **0.4215** | 0.3779 | 0.5858 |
| 35 | 0.1813 | **0.2148** | *0.0552* | 0.0665 | 0.0617 | 0.0174 | **0.1755** |
| 38 | 0.1249 | 0.1621 | 0.1035 | ***0.0665*** | **0.1500** | **0.1321** | 0.1097 |
| 39 | 0.4086 | **0.3235** | 0.4141 | 0.1329 | 0.5293 | ***0.6085*** | 1.0000 |
| 40 | *0.4505* | 0.1931 | **0.2070** | 0.1661 | 0.4126 | 0.3260 | **0.1755** |
| 42 | 0.2416 | 0.2875 | 0.3278 | 0.1329 | **0.1952** | **0.2388** | *0.2627* |

The effectiveness of this approach is evaluated by assuming varying percentage of data as missing, starting from 10% to 30%. The data assumed as missing is Missing Completely At Random (MCAR). MCAR implies that data assumed as missing does not follow any pattern with respect to data present or missing. All the values assumed as missing have equal probability to be identified as missing, Little and Rubin (2002).

The values assumed as missing in 10% missing case are shown as italics in Table 7.3. The values assumed as missing in 20% missing case are underlined and shown in Table 7.3. The values assumed as missing in 30% missing case are represented in bold and shown in Table 7.3. If a value is both italic and underlined then it is assumed as missing in both 10% and 20% missing cases. Similarly, if the value is both italic and bold then it is assumed as missing in both 10% and 30% missing cases.

The number of values assumed as missing in case of 10%, 20%, and 30% are 16, 32, and 48 respectively. In each case the number of missing input values equals to number of missing output values. For each case the difference between the lowest and the highest of the assumed missing values are calculated and this value is known as range. Range for 10%, 20%, and 30% missing cases are shown in Table 7.4. The range demonstrates the variability of the missing data, higher the value greater the difficulty to recover the missing data precisely.

**Table 7-4: Number of Missing Values**

| No. | Title | Range | # of Missing Values | # of Missing Inputs | # of Missing Outputs |
|---|---|---|---|---|---|
| 1 | 10% Missing | 0.9603 | 16 | 8 | 8 |
| 2 | 20% Missing | 0.9675 | 32 | 16 | 16 |
| 3 | 30% Missing | 0.7517 | 48 | 24 | 24 |

### *7.3.1 Relative Intervals*

In this case the intervals are relatively constructed for missing values based on the actual values (assumed as missing), rest of the values remain crisp. Intervals constructed equals 50% times of the actual value. Where the lower bound is 25% times less than the actual value and upper bound is 25% times greater than the actual value. Hence, the interval range constructed equals 50% of the actual value, varying symmetrically at ∓25% around the center point (the actual known value). This implies that when alpha equals to 0.5 the crisp values of the interval range represent the actual missing values. Based on a common alpha value for all the interval ranges, crisp values are determined using linear interpolations. Thus the relative interval around the central point $x_i$ is calculated as:

$$x_L = max\begin{cases} x_i - 0.25 * x_i \\ 0 \end{cases} \quad and \quad x_U = min\begin{cases} x_i + 0.25 * x_i \\ 1 \end{cases}$$

### *7.3.2 Fixed Intervals*

In this case also the intervals are constructed for missing values based on the actual values (assumed as missing), rest of the values remain crisp. Intervals are fixed around the actual values. Hence, the intervals are symmetrically varying at ∓0.25 around the center point (the actual known value). Based on a common alpha value for all the interval ranges, crisp values are determined using linear interpolations. These calculated crisp values are substituted back into the dataset to carry out the DEA analysis and to determine the efficiency scores. Thus the fixed interval around the central point $x_i$ is calculated as:

$$x_L = max\begin{cases} x_i - 0.25 \\ 0 \end{cases} \quad and \quad x_U = min\begin{cases} x_i + 0.25 \\ 1 \end{cases}$$

For more details on construction of the relative and fixed intervals refer to Appendix D.

## 7.4 Results and Discussions

The efficiency scores are determined for different alpha values based on the interval range constructed, for both relative and fixed intervals. The efficiency scores of the clinics are determined based on CCR Input oriented model. The efficiency scores obtained by varying the percentage of missing data from 10% to 30% for relative intervals are shown in Tables 7.5 to 7.7

respectively for different values of alpha. The efficiency scores obtained by varying the percentage of missing data from 10% to 30% for fixed intervals are shown in Tables 7.8 to 7.10 respectively for different values of alpha. Efficiency scores of all the 22 DMUs obtained for a particular alpha value are summed. The alpha value corresponding to the highest summation value of efficiency scores is chosen as the best. Hence the best alpha value chosen bestows most of the DMUs with best efficiency scores. The best chosen value of alpha and the corresponding efficiency score for the 22 clinics are highlighted in each case.

**Table 7-5: Efficiency Scores for different Alpha Values, 10% Missing, Relative Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.600 | 0.591 | 0.582 | 0.574 | 0.566 | 0.558 | 0.551 | 0.543 | 0.536 | 0.529 | 0.523 |
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 |
| 11 | 0.698 | 0.675 | 0.657 | 0.641 | 0.623 | 0.611 | 0.600 | 0.591 | 0.579 | 0.569 | 0.559 |
| 12 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 14 | 0.752 | 0.730 | 0.709 | 0.690 | 0.672 | 0.655 | 0.639 | 0.625 | 0.612 | 0.608 | 0.604 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 0.667 | 0.659 | 0.651 | 0.643 | 0.635 | 0.629 | 0.622 | 0.616 | 0.610 | 0.605 | 0.599 |
| 17 | 1.000 | 0.988 | 0.973 | 0.959 | 0.946 | 0.933 | 0.920 | 0.908 | 0.896 | 0.884 | 0.873 |
| 20 | 0.650 | 0.639 | 0.633 | 0.627 | 0.622 | 0.617 | 0.612 | 0.608 | 0.603 | 0.599 | 0.595 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 0.345 | 0.344 | 0.344 | 0.343 | 0.343 | 0.342 | 0.342 | 0.342 | 0.341 | 0.341 | 0.341 |
| 29 | 0.637 | 0.628 | 0.620 | 0.612 | 0.605 | 0.598 | 0.591 | 0.585 | 0.578 | 0.572 | 0.566 |
| 33 | 0.788 | 0.797 | 0.806 | 0.815 | 0.825 | 0.835 | 0.846 | 0.856 | 0.867 | 0.878 | 0.889 |
| 34 | 0.452 | 0.446 | 0.440 | 0.435 | 0.430 | 0.426 | 0.421 | 0.417 | 0.413 | 0.409 | 0.406 |
| 35 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 0.974 | 0.950 | 0.928 | 0.907 |
| 38 | 0.973 | 0.967 | 0.959 | 0.950 | 0.941 | 0.932 | 0.924 | 0.916 | 0.909 | 0.902 | 0.896 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 42 | 0.650 | 0.642 | 0.634 | 0.626 | 0.618 | 0.610 | 0.602 | 0.594 | 0.585 | 0.577 | 0.568 |
| **SUM** | **18.072** | **17.965** | **17.868** | **17.776** | **17.686** | **17.605** | **17.529** | **17.433** | **17.342** | **17.261** | **17.184** |

**Table 7-6: Efficiency Scores for different Alpha Values, 20% Missing, Relative Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.451 | 0.476 | 0.501 | 0.521 | 0.541 | 0.558 | 0.575 | 0.593 | 0.610 | 0.627 | 0.644 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 0.645 | 0.688 | 0.731 | 0.774 | 0.817 | 0.860 | 0.903 | 0.946 | 0.989 | 1.000 | 1.000 |
| 11 | 0.573 | 0.578 | 0.588 | 0.600 | 0.602 | 0.611 | 0.625 | 0.640 | 0.656 | 0.674 | 0.693 |
| 12 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 14 | 0.450 | 0.490 | 0.532 | 0.577 | 0.617 | 0.655 | 0.692 | 0.729 | 0.777 | 0.833 | 0.893 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 0.621 | 0.622 | 0.623 | 0.625 | 0.627 | 0.629 | 0.630 | 0.636 | 0.653 | 0.671 | 0.689 |
| 17 | 0.841 | 0.858 | 0.876 | 0.894 | 0.913 | 0.933 | 0.953 | 0.974 | 0.990 | 1.000 | 1.000 |
| 20 | 0.566 | 0.577 | 0.593 | 0.602 | 0.609 | 0.617 | 0.624 | 0.631 | 0.639 | 0.645 | 0.650 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 0.376 | 0.359 | 0.346 | 0.335 | 0.333 | 0.342 | 0.354 | 0.366 | 0.378 | 0.386 | 0.393 |
| 29 | 0.493 | 0.524 | 0.552 | 0.568 | 0.583 | 0.598 | 0.612 | 0.627 | 0.641 | 0.661 | 0.682 |
| 33 | 0.945 | 0.919 | 0.892 | 0.871 | 0.844 | 0.835 | 0.833 | 0.831 | 0.829 | 0.824 | 0.825 |
| 34 | 0.365 | 0.379 | 0.393 | 0.405 | 0.416 | 0.426 | 0.435 | 0.444 | 0.453 | 0.463 | 0.472 |
| 35 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.994 | 0.965 | 0.938 | 0.913 | 0.890 |
| 38 | 0.761 | 0.802 | 0.852 | 0.881 | 0.907 | 0.932 | 0.955 | 0.977 | 0.996 | 1.000 | 1.000 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 42 | 0.597 | 0.595 | 0.606 | 0.614 | 0.610 | 0.610 | 0.611 | 0.612 | 0.611 | 0.611 | 0.614 |
| **SUM** | **16.686** | **16.866** | **17.084** | **17.267** | **17.420** | **17.605** | **17.798** | **17.969** | **18.160** | **18.307** | **18.445** |

**Table 7-7: Efficiency Scores for different Alpha Values, 30% Missing, Relative Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.499 | 0.500 | 0.515 | 0.529 | 0.541 | 0.558 | 0.579 | 0.602 | 0.628 | 0.656 | 0.686 |
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 1.000 | 1.000 | 1.000 | 0.950 | 0.902 | 0.860 | 0.821 | 0.786 | 0.754 | 0.724 | 0.697 |
| 11 | 0.779 | 0.750 | 0.706 | 0.667 | 0.632 | 0.610 | 0.591 | 0.576 | 0.567 | 0.582 | 0.598 |
| 12 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 14 | 0.757 | 0.720 | 0.692 | 0.667 | 0.645 | 0.655 | 0.680 | 0.701 | 0.717 | 0.729 | 0.737 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 0.604 | 0.615 | 0.626 | 0.632 | 0.633 | 0.629 | 0.622 | 0.627 | 0.641 | 0.663 | 0.685 |
| 17 | 0.964 | 0.958 | 0.952 | 0.946 | 0.939 | 0.933 | 0.926 | 0.920 | 0.913 | 0.907 | 0.900 |
| 20 | 0.630 | 0.637 | 0.638 | 0.630 | 0.623 | 0.617 | 0.612 | 0.608 | 0.610 | 0.612 | 0.610 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 0.431 | 0.409 | 0.388 | 0.370 | 0.354 | 0.340 | 0.327 | 0.318 | 0.317 | 0.315 | 0.313 |
| 29 | 0.595 | 0.601 | 0.613 | 0.617 | 0.605 | 0.598 | 0.592 | 0.588 | 0.585 | 0.583 | 0.582 |
| 33 | 0.863 | 0.869 | 0.871 | 0.864 | 0.852 | 0.835 | 0.823 | 0.828 | 0.832 | 0.829 | 0.814 |
| 34 | 0.539 | 0.498 | 0.472 | 0.454 | 0.437 | 0.426 | 0.420 | 0.427 | 0.435 | 0.444 | 0.453 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 35 | 0.769 | 0.820 | 0.871 | 0.923 | 0.974 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 38 | 0.742 | 0.787 | 0.829 | 0.863 | 0.899 | 0.931 | 0.961 | 0.992 | 1.000 | 1.000 | 1.000 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 42 | 0.530 | 0.538 | 0.550 | 0.570 | 0.591 | 0.610 | 0.627 | 0.640 | 0.653 | 0.675 | 0.698 |
| **SUM** | **17.702** | **17.703** | **17.725** | **17.680** | **17.628** | **17.601** | **17.581** | **17.613** | **17.651** | **17.718** | **17.773** |

**Table 7-8: Efficiency Scores for different Alpha Values, 10% Missing, Fixed Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.954 | 0.774 | 0.658 | 0.601 | 0.530 | 0.489 | 0.465 | 0.447 | 0.427 | 0.406 | 0.386 |
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 0.682 | 0.824 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 | 0.860 |
| 11 | 1.000 | 0.943 | 0.831 | 0.741 | 0.662 | 0.611 | 0.568 | 0.525 | 0.484 | 0.447 | 0.412 |
| 12 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.967 |
| 14 | 1.000 | 1.000 | 1.000 | 0.698 | 0.590 | 0.550 | 0.530 | 0.520 | 0.520 | 0.520 | 0.520 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 0.876 | 0.859 | 0.772 | 0.690 | 0.623 | 0.578 | 0.553 | 0.529 | 0.508 | 0.490 | 0.474 |
| 17 | 0.968 | 1.000 | 1.000 | 1.000 | 0.939 | 0.826 | 0.737 | 0.673 | 0.627 | 0.593 | 0.564 |
| 20 | 0.581 | 0.569 | 0.631 | 0.611 | 0.574 | 0.528 | 0.481 | 0.469 | 0.469 | 0.469 | 0.469 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 0.370 | 0.359 | 0.356 | 0.347 | 0.341 | 0.339 | 0.337 | 0.337 | 0.336 | 0.336 | 0.335 |
| 29 | 0.918 | 1.000 | 0.750 | 0.637 | 0.578 | 0.541 | 0.506 | 0.473 | 0.442 | 0.413 | 0.391 |
| 33 | 0.444 | 0.551 | 0.678 | 0.707 | 0.698 | 0.692 | 0.749 | 0.807 | 0.864 | 0.922 | 0.979 |
| 34 | 0.509 | 0.491 | 0.477 | 0.458 | 0.428 | 0.399 | 0.380 | 0.361 | 0.343 | 0.326 | 0.308 |
| 35 | 1.000 | 1.000 | 1.000 | 0.747 | 0.565 | 0.459 | 0.392 | 0.388 | 0.387 | 0.385 | 0.383 |
| 38 | 1.000 | 1.000 | 0.904 | 0.943 | 0.920 | 0.874 | 0.818 | 0.760 | 0.702 | 0.658 | 0.658 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.991 | 0.921 | 0.921 | 0.921 |
| 42 | 0.565 | 0.623 | 0.655 | 0.639 | 0.622 | 0.599 | 0.579 | 0.584 | 0.620 | 0.679 | 0.736 |
| **SUM** | **18.869** | **18.993** | **18.574** | **17.677** | **16.931** | **16.344** | **15.955** | **15.724** | **15.509** | **15.424** | **15.362** |

**Table 7-9: Efficiency Scores for different Alpha Values, 20% Missing, Fixed Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.900 | 0.810 |
| 4 | 0.365 | 0.377 | 0.438 | 0.480 | 0.516 | 0.569 | 0.681 | 0.736 | 0.731 | 0.726 | 0.725 |
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 0.471 | 0.522 | 0.659 | 0.861 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 11 | 0.453 | 0.495 | 0.532 | 0.565 | 0.600 | 0.619 | 0.658 | 0.703 | 0.753 | 0.788 | 0.820 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 12 | 0.674 | 0.857 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 14 | 0.288 | 0.330 | 0.413 | 0.576 | 0.693 | 0.901 | 0.906 | 0.910 | 0.913 | 0.915 | 0.917 |
| 15 | 0.680 | 0.893 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 0.557 | 0.607 | 0.618 | 0.687 | 0.735 | 0.846 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 17 | 0.716 | 0.804 | 0.984 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 20 | 0.506 | 0.541 | 0.567 | 0.587 | 0.635 | 0.666 | 0.671 | 0.657 | 0.627 | 0.597 | 0.573 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 0.921 | 0.843 | 0.715 | 0.563 | 0.494 | 0.496 | 0.519 | 0.566 | 0.643 | 0.723 | 0.803 |
| 29 | 0.369 | 0.383 | 0.418 | 0.468 | 0.522 | 0.627 | 0.770 | 0.830 | 0.857 | 0.868 | 0.897 |
| 33 | 1.000 | 1.000 | 1.000 | 0.963 | 0.916 | 0.925 | 0.855 | 0.770 | 0.712 | 0.670 | 0.636 |
| 34 | 0.281 | 0.307 | 0.372 | 0.403 | 0.403 | 0.427 | 0.450 | 0.474 | 0.521 | 0.571 | 0.620 |
| 35 | 0.843 | 1.000 | 0.928 | 0.694 | 0.563 | 0.506 | 0.607 | 0.708 | 0.809 | 0.759 | 0.685 |
| 38 | 1.000 | 0.896 | 0.734 | 0.862 | 0.999 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.953 |
| 42 | 0.882 | 0.790 | 0.711 | 0.662 | 0.620 | 0.612 | 0.637 | 0.743 | 0.951 | 1.000 | 1.000 |
| **SUM** | **16.006** | **16.646** | **17.088** | **17.370** | **17.698** | **18.193** | **18.753** | **19.098** | **19.516** | **19.515** | **19.437** |

**Table 7-10: Efficiency Scores for different Alpha Values, 30% Missing, Fixed Intervals**

| **Key No** | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 4 | 0.481 | 0.485 | 0.474 | 0.493 | 0.545 | 0.643 | 0.688 | 0.747 | 0.812 | 0.851 | 0.850 |
| 5 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 7 | 1.000 | 1.000 | 1.000 | 0.772 | 0.651 | 0.719 | 0.722 | 0.707 | 0.746 | 0.849 | 0.960 |
| 11 | 0.582 | 0.680 | 0.703 | 0.668 | 0.634 | 0.611 | 0.585 | 0.542 | 0.508 | 0.509 | 0.514 |
| 12 | 0.600 | 1.000 | 1.000 | 0.918 | 0.931 | 0.866 | 0.868 | 0.879 | 0.919 | 0.971 | 0.980 |
| 13 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.982 | 0.902 | 0.833 |
| 14 | 1.000 | 1.000 | 0.938 | 0.761 | 0.889 | 0.947 | 0.830 | 0.756 | 0.731 | 0.736 | 0.739 |
| 15 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 16 | 1.000 | 0.694 | 0.621 | 0.628 | 0.777 | 0.993 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 17 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 20 | 0.519 | 0.543 | 0.568 | 0.588 | 0.603 | 0.639 | 0.654 | 0.630 | 0.610 | 0.572 | 0.549 |
| 22 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 23 | 1.000 | 1.000 | 0.769 | 0.542 | 0.456 | 0.444 | 0.401 | 0.376 | 0.364 | 0.357 | 0.349 |
| 29 | 1.000 | 0.871 | 0.680 | 0.601 | 0.677 | 0.664 | 0.623 | 0.625 | 0.643 | 0.679 | 0.726 |
| 33 | 0.631 | 0.840 | 0.862 | 0.844 | 0.793 | 0.930 | 0.922 | 0.884 | 0.854 | 0.775 | 0.709 |
| 34 | 0.611 | 0.601 | 0.539 | 0.472 | 0.443 | 0.447 | 0.476 | 0.490 | 0.503 | 0.522 | 0.541 |
| 35 | 1.000 | 0.457 | 0.580 | 0.746 | 0.994 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 38 | 1.000 | 0.461 | 0.550 | 0.792 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |
| 39 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 |

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 40 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.971 | 0.880 | 0.917 | 0.996 | 1.000 |
| 42 | 0.384 | 0.413 | 0.439 | 0.469 | 0.576 | 0.674 | 0.782 | 0.894 | 1.000 | 1.000 | 1.000 |
| **SUM** | **18.807** | **18.046** | **17.722** | **17.294** | **17.967** | **18.579** | **18.522** | **18.410** | **18.588** | **18.719** | **18.750** |

The effectiveness of the recovered efficiency scores are evaluated based on mean absolute difference between the actual and recovered efficiency scores. The Mean Absolute Deviation values for different cases are shown in Table 7.11.

**Table 7-11: Mean Absolute Deviation for recovered efficiency scores**

| | **Relative Intervals** | | | **Fixed Intervals** | | |
|---|---|---|---|---|---|---|
| | **10%** | **20%** | **30%** | **10%** | **20%** | **30%** |
| **Mean Absolute Deviation** | 0.026 | 0.050 | 0.033 | 0.096 | 0.115 | 0.148 |

### *7.4.1 Effects of Interval Approach on DEA Results*

In the previous section we assumed varying quantities of data as missing starting from 10% to 30% and determined the efficiency scores for different values of alpha. The best value of alpha is chosen based on summation of the efficiency scores for all DMUs. The summation of efficiency scores as a function of alpha for relative intervals considered in the above section is shown in Figure 7.1. The summation of efficiency scores as a function of alpha for fixed intervals considered in the above section is shown in Figure 7.2.

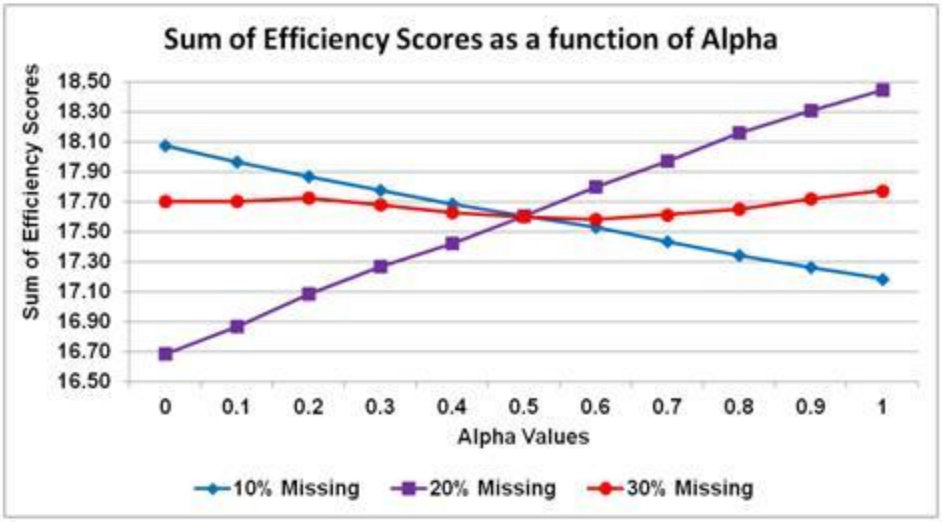

**Figure 7-1: Summation of Efficiency Scores as a function of Alpha, Relative Intervals**

Observation of Figure 7.1 shows that the nature and range (difference between the highest and lowest) of the efficiency scores does not really depend on the percentage of missing values. The range of the DEA efficiency scores in cases of 20% missing is greater when compared to 10% and 10% be greater when compared to 30% missing. Moreover, the nature of the DEA efficiency scores also does not depend on the amount of missing values. The DEA efficiency scores in case of 10% missing values are linearly decreasing, increasing linearly in case of 20% missing, and in case of 30% missing it follows both increasing and decreasing nature.

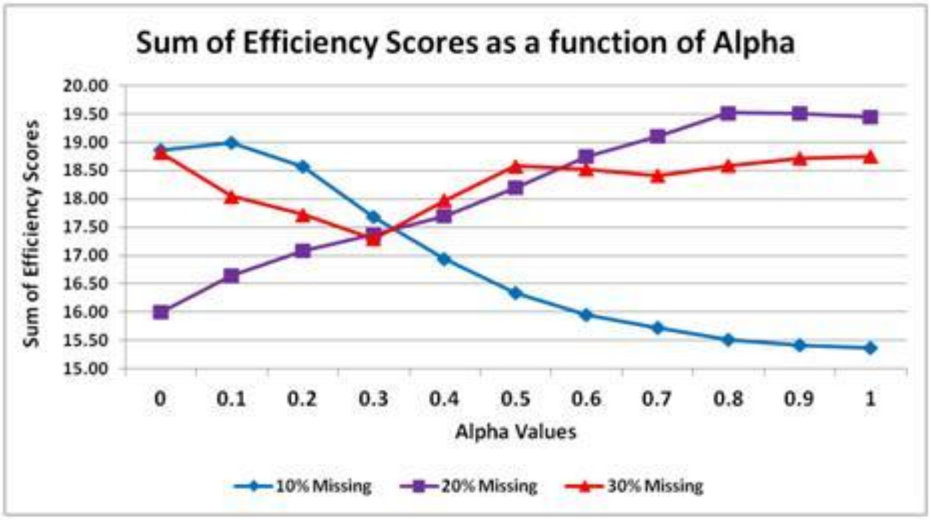

**Figure 7-2: Summation of Efficiency Scores as a function of Alpha, Fixed Intervals**

Similarly we can also observe from Figure 7.2 that the nature and range (difference between the highest and lowest) of the efficiency scores does not really depend on the percentage of missing values.

Based on the Mean Absolute Deviation values shown in Table 7.11, we can identify that recovered efficiency scores are less deviated from the actual scores in case of relative intervals. This means that outcomes of DEA are directly proportional to the accuracy of constructing the narrow interval ranges.

## 7.5 Conclusions

Generally in case of missing values being estimated within an interval ranges there exists multiple efficiency scores leaving the decision to the individuals. This chapter focuses on determining the best crisp DEA efficiency score out of interval ranges created to estimate the missing values.

This chapter provides literature review of interval based DEA models to handle the issue of missing values. This chapter proposes new approach based on interval range to determine the best crisp efficiency scores. The missing values in the dataset are estimated with in interval range based on the actual values. Interval range is broken down into crisp values based on linear interpolations. This chapter considers the concept of common alpha for all the interval ranges in order to avoid the fuzzy environment, and to reduce the total number of computations by 50%. Crisp values are replaced back into the dataset to carry out the Data Envelopment Analysis. Best value of alpha is chosen which provides the best efficiency scores to most of the DMUs.

The interval based approach is illustrated using a complete data set of the 22 clinics, with varying levels of assumed missing values, ranging from 10% to 30%. The values assumed as missing are considered to be Missing Completely At Random (MCAR). The values assumed as missing are estimated within an interval range. Interval ranges are broken down into crisp values and are replaced back into the dataset to execute the DEA. Efficiency scores are determined for the different values of alpha. Based on the summation of all the efficiency score, best value of alpha is chosen. Best alpha value chosen bestows most of the DMUs with best possible efficiency scores. Two types of interval ranges are constructed to show that accuracy of the DEA outcomes depend on the construction of intervals.

The summation of efficiency scores as a function of missing data is graphically illustrated. Based on this illustration we can suggest that percentage of missing values cannot influence the range and nature of the efficiency scores. The results suggest that the efficiency scores are fairly insensitive to the missing data.

As a summary, this chapter provides an effective and practical approach for replacing missing values needed for a DEA analysis based on interval values.

# Chapter 8 - SUMMARY AND FUTURE RESEARCH

This chapter provides brief summary of the research work carried out, its outcomes and as well as the scope for future research in this area.

This thesis proposes new methodologies based on three different platforms such as correlation, clustering, and interval approach to execute Data Envelopment Analysis (DEA) with sparse data in an effective manner. It also provides the motivation and necessity for scrutinizing the proposed research. The objective is to evaluate the productivity of 41 member clinics of Kansas Association of Medically Underserved (KAMU) with sparse data. In order to achieve this primary focal point, our goal is to develop new methods to determine the missing values and then to execute DEA in a reliable manner.

This thesis provides a thorough background to Data Envelopment Analysis, clearly states the different concepts germane to this research, conducts an exhaustive literature review to analyze the importance of efficiency measurement techniques in the field of healthcare, also provides the summary of missing data treatment methods to handle DEA, and identified limitations associated with few of these methods to improve their effectiveness. It also presents the guidelines to prepare the data by identifying the issues, providing the measures to clean the data and to perform DEA more effectively.

This thesis significantly contributed few new methodologies to this area named as Average Ratio Method, and Cluster Dispersion. It successfully incorporated clustering methodology as an intermediate approach to determine the missing values for DEA and also studied the influence of such recovered values on the efficiency score results of DEA. This thesis identified productive ways to determine the crisp efficiency scores from interval values estimated by experts. The effectiveness of these proposed methods are evaluated by comparing few of them by the data acquired from literature works, and others by their capability to determine the assumed missing values within the proximity of actual values. These proposed methods are tested for different levels of missing values, up to 40% of the data assumed as missing. They are also tested for different possible nature of missing values. These methods can serve as benchmarks in this area, to recover missing data. These methods can generate the efficiency scores within close proximity to the actual efficiency scores.

First method named as Average Ratio Method (ARM) is based on the concept of correlation between two variables. Accuracy of this methodology depends on the level of correlation between them. The effectiveness of this method is way more when compared to other basic methods, such as case-wise deletion and replacing the missing values by average.

Second method is based on the concept of modified fuzzy c-means clustering algorithm which can handle missing values, is an existing algorithm. Identified the two primary issues associated with this method, suggested approaches to eliminate these issues and improved the effectiveness of this method further. Identified the convergence failure of existing Cluster Dispersion algorithm, illustrated the reasons. Three new approaches are proposed based on the similar lines of existing cluster dispersion methodology.

Third method is based on interval approach to handle the issue of missing values. Missing values are replaced by interval ranges estimated by experts. Crisp efficiency scores are identified in similar lines to how DEA determines efficiency scores using the best set of weights. Identification of crisp efficiency scores out of interval values is something uncommon in this area. The concept of common alpha value reduces the total number of computations by 50% when compared to other methods.

As a summary, this thesis provides valuable methods to recover the missing values, and evaluated the effectiveness of these methods. This thesis also provides guiding principles for someone looking into the practical approach for executing Data Envelopment Analysis with sparse data.

## 8.1 Future Research

For future research in this area, one can look into the core aspect of the Data Envelopment Analysis, identifying the linear programming structure to estimate the missing values. This can be achieved in two ways, one by controlling the behavior of weights associated with the missing values. The other can be achieved by using the sensitivity analysis concept of linear programming methodology to estimate the range for missing value where the DMU continues to be efficient for that interval range. Converse to the core aspect, one can look into other domains to determine the missing values (similar to clustering approach in this thesis) and using it as an intermediate approach. Then evaluate the effectiveness and influence of these intermediate methods on DEA outcomes.

# References

Andersen, P., and Petersen, N.C. (1993). A Procedure for Ranking Efficient Units in Data Envelopment Analysis. Management Science, 39, 1261-1264.

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(9), 1078-1092.

Banker, R.D., Conrad, R.F., and Strauss, R.P. (1986). A Comparative Application of Data Envelopment Analysis and Translog Methods: An Illustrative Study of Hospital Production. Management Science, 32(1), 30-44.

Banker, R.D., and Morey, R. (1986a). Efficiency Analysis for Exogenously Fixed Inputs and Outputs. Operations Research, 34(4), 513-521.

Banker, R.D., and Morey, R. (1986b). The Use of Categorical Variables in Data Envelopment Analysis. Management Science, 32(12), 1613-1627.

Banker, R.D., Charnes, A., Cooper, W.W., Swarts, J., and Thomas, D. (1989). An Introduction to Data Envelopment Analysis with Some of its Models and Their Uses. Research in Government and Nonprofit Accounting, 5, 125-163.

Basson, M.D., and Butler, T. (2006). Evaluation of Operating Room Suite Efficiency in the Veterans Health Administration System by Using Data Envelopment Analysis. The American Journal of Surgery, 192(5), 649-656.

Becker D., 2011. DEA Bib. Retrieved October, 2011, from http://www.deabib.org/journals.html#x6-160005.

Benson, H. R. (1994). An Introduction to Benchmarking in Healthcare. Radiology Management, 16(4), 35-39.

Berger, A.N., and Humphery, D.B. (1997). Efficiency of Financial Institutions: International Survey and Directions for Future Research. Financial Institutions Center Working Paper 97-05, The Wharton School, University of Pennsylvania.

Björkgren, M.A., Häkkinen, U., and Linna, M. (2001). Measuring Efficiency of Long-Term Care Units in Finland. Health Care Management Science, 4(3), 193-200.

Blank, J.L.T., and Valdmanis, V.G. (2010). Environmental Factors and Productivity on Dutch Hospitals: A Semi-Parametric Approach. Health Care Management Science, 13(1), 27-34.

Bryce, C.J., Engberg, J.B., and Wholey, D.R. (2000). Comparing the Agreement among Alternative Models in Evaluating HMO Efficiency. Health Services Research, 35, 509-528.

Centers for Medicare & Medicaid Services. Retrieved September, 2011, from https://www.cms.gov/nationalhealthexpenddata/02_nationalhealthaccountshistorical.asp.

Chacon, M., Luci, O. (2003). Patients Classification by Risk using Cluster Analysis And Genetic Algorithms. Progress in Pattern Recognition, Speech and Image Analysis. 8th Iberoamerican Congress on Pattern Recognition, CIARP. Proceedings (Lecture Notes in Computer Science. 2905), 350-358.

Charnes, A., Cooper, W.W., and Rhodes, E. (1978). Measuring the Efficiency of Decision Making Units. European Journal of Operations Research, 2, 429-444.

Charnes, A., Cooper, W.W., Golany, B., Seiford, L.M., and Stutz, J. (1985). Foundations of Data Envelopment Analysis for Pareto-Koopmans Efficient Empirical Production Functions. Journal of Econometrics, 30, 91-107.

Charnes, A., Cooper, W.W., and Huang, Z.M. (1990). Polyhedral Cone-Ratio DEA Models with an Illustrative Application to Large Commercial Banks. Journal of Econometrics, 46, 73-91.

Chen, Y., and Zhu, J. (2007). Chapter 3: Interval and Ordinal Data. In Zhu, J., and Cook, W.D. (Ed.), Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis (pp. 35-62). New York: Springer Science.

Chilingerian, J.A., and Sherman H.D. (2004). Chapter 17: Health Care Applications. In Cooper, W.W., Seiford, L.M., and Zhu, J., Handbook on Data Envelopment Analysis (pp. 481-538). Boston: Kluwer Academic Publishers.

Coelli, T.J., Rao, D.S.P., O'Donnell, C.J., Battese, G.E. (2005). An Introduction to Efficiency and Productivity Analysis. New York: Springer Science.

Congdon, P. (1997). Multilevel and Clustering Analysis of Health Outcomes in Small Areas. European Journal of Population, 13(4), 305-338.Cook, W.D., Green, R.H., and Zhu, J. (2006). Dual-role factors in DEA. IIE Transactions, 38(2), 105-115.

Cook, W.D., and Zhu, J. (2007). Classifying Inputs and Outputs in Data Envelopment Analysis. European Journal of Operations Research, 180, 692-699.

Cook, W.D., and Zhu, J. (2007). Chapter 2: Rank Order Data in DEA. In Zhu, J., and Cook, W.D. (Ed.), Modeling Data Irregularities And Structural Complexities In Data Envelopment Analysis (pp. 13-34). New York: Springer Science.

Cooper, W.W., Seiford, L.M., and Zhu, J. (2004). Chapter 1: Data Envelopment Analysis: History, Models and Interpretations. In Cooper, W.W., Seiford, L.M., and Zhu, J., Handbook on Data Envelopment Analysis (pp. 481-538). Boston: Kluwer Academic Publisher.

Cooper, W.W., Seiford, L.M., Tone, K. (2007). Data Envelopment Analysis: A Comprehensive Text with Models, Application, References and DEA Solver Software. New York: Springer Science.

Despotis, D.K., and Smirlis, Y. (2002). Data Envelopment Analysis with Imprecise Data. European Journal of Operational Research, 140, 24-36.

Dunn, J.C. (1973). A Fuzzy Relative of the ISODATA Process and Its use in Detecting Compact Well-Separated Clusters. Journal of Cybernetics, 3(3), 32-57.

Dyson, R.G., and Thanassoulis, E. (1988). Reducing Weight Flexibility in Data Envelopment Analysis. Journal of Operational Research Society, 39(6), 563-576.

Edwards, A., and Cavalli-Sforza, L. (1965). A Method for Cluster Analysis. Biometrics, 21(2), 362-375.

Evidence on the Costs and Benefits of Health Information Technology. Retrieved September, 2011, from http://www.cbo.gov/ftpdocs/91xx/doc9168/MainText.3.1.shtml.

Fare, R., and Grosskopf, S. (2002). Two Perspectives on DEA: Unveiling the link between CCR and Shephard. Journal of Productivity Analysis, 17(1-2), 41-47.

Faris, P.D., Ghali, W.A., Brant, R., Norris, C.M., Galbraith, P.D., and Knudtson, M.L. (2002). Multiple Imputation versus Data Enhancement for Dealing with Missing Data in Observational Health Care Outcome Analyses. Journal of Clinical Epidemiology, 55(2), 184-191.

Farrell, M. J. (1957). The Measurement of Productive Efficiency. Journal of Royal Statistical Society, 120(11), 254-290.

Florek, K., Lukaszewicz, J., Steinhaus, H., and Zubrzycki, S. (1951). Sur la liaison et la division des points d'un ensemble fini. Colloquium Mathematicum, 2, 282–285.

Fujikawa, Y., and Ho, T. (2002). Cluster-based algorithms for dealing with missing values. In Cheng, M.-S., Yu, P. S., and Liu, B., editors, “Advances in Knowledge Discovery and Data Mining”, Proceedings of the 6th Pacific-Asia Conference, PAKDD Taipei, Taiwan, volume 2336 of Lecture Notes in Computer Science, 549–554, New York.

Gan, G., Ma, C., and Wu, J. (2007). Data Clustering: Theory, Algorithms, and Applications. ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia.

Garavaglia, G., Lettieri, E., Agasisti, T., and Lopez, S. (2011). Efficiency and quality of care in nursing homes: an Italian case study. Health Care Management Science, 14(1), 22-35.

Graham, J. W. (2009). Missing Data Analysis: Making it Work in the Real World. Annual Review of Psychology, 60, 549-576.

Hatefi, S.M., Jolai, F., and Kor, H. (2009). A New Model for Classifying Inputs and Outputs and Evaluating the DMUs Efficiency in DEA Based on Cobb-Douglas Production Function. International Conference on Computer Engineering and Technology, 1, 390-394.

Hathaway, R.J., and Bezdek, J.C. (2001). Fuzzy C-Means Clustering of Incomplete Data. IEEE Transactions on Systems, Man, and Cybernetics-part b. Cybernetics, 31(5), 735-744.

Hathaway, R.J., Hu, Y., and Bezdek, J.C. (2001). Local Convergence of Tri-Level Alternating Optimization. Neural, Parallel, and Scientific Computation, 9, 19-28.

Helmig, B., and Lapsley, I. (2001). On the Efficiency of Public, Welfare and Private Hospitals in Germany over Time: A Sectoral Data Envelopment Analysis Study. Health Services Management Research, 14(4), 263-274.

Himmelspach, L., and Conrad, S. (2010). Fuzzy Clustering of Incomplete Data Based on Cluster Dispersion. Computational Intelligence for Knowledge-Based Systems Design. 6178, 59-68.

Hofmarcher, M.M., Paterson, L., and Riedel, M. (2002). Measuring Hospital Efficiency in Austria-A DEA Approach. Health Care Management Science, 5(1), 7-14.

Hollingsworth, B., Dawson, P.J., and Maniadakis, N. (1999). Efficiency Measurement of Health Care: A Review of Non-Parametric Methods and Applications. Health Care Management Science, 2, 161-172.

Hollingsworth , B., and Parkin, D. (2001). The Efficiency of the Delivery of Neonatal Care in the UK. Journal of Public Health Medicine, 23(1), 47-50.

Hollingsworth, B. (2003). Non-Parametric and Parametric Applications Measuring Efficiency in Health Care. Health Care Management Science, 6, 203-218.

Jacobs , R. (2001). Alternative Methods to Examine Hospital Efficiency: Data Envelopment Analysis and Stochastic Frontier Analysis. Health Care Management Science, 4, 103-115.

Jain, A., and Dubes, R. (1988). Algorithms for Clustering Data. New Jersey: Prentice–Hall Englewood Cliffs.

Johnson, S. (1967). Hierarchical Clustering Schemes. Psychometrika, 32(3), 241-254.

Kansas Association of Medically Underserved (KAMU) Mission. Retrieved September, 2011, from http://www.kamuonline.org/about.php.

Kao, C., and Liu, S.T. (2000). Data Envelopment Analysis with Missing Data: An application to University Libraries in Taiwan. Journal of Operational Research Society, 51 (8), 897–905.

Kao, C., and Liu, S.T. (2007). Chapter 16: Data Envelopment Analysis with Missing Data. In Zhu, J., and Cook, W.D. (Ed.), Modeling Data Irregularities and Structural Complexities in Data Envelopment Analysis (pp. 291-304). New York: Springer Science.

Kaufman, L., and Rousseeuw, P. (1990). Finding Groups in Data-An Introduction to Cluster Analysis. New York: Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons.

Kirch, W. (2008). Encyclopedia of Public Health (pp. 523). New York: Springer Science.

Kuosmanen, T. (2002). Modeling Blank Data Entries in Data Envelopment Analysis. Econometrics working paper archive at WUSTL, No. 0210001.

Kuosmanen, T. (2009). Data Envelopment Analysis with Missing Data. Journal of the Operational Research Society, 60, 1767-1774.

Little, R.J.A., and Rubin, D.B. (2002). Statistical Analysis with Missing Data. New York: John Wiley & Sons.

MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, University of California Press, 281–297.

Mukherjee, K., Santerre, R., and Zhang, N.J. (2010). Explaining the Efficiency of Local Health Departments in the U.S.: An Exploratory Analysis. Health Care Management Science, 13(4), 378-387.

Nathanson, B.H., Higgins, T.L., Giglio, R.J., Munshi, I.A., and Steingrub, J.S. (2003). An Exploratory Study Using Data Envelopment Analysis to Assess Neurotrauma Patients in the Intensive Care Unit. Health Care Management Science, 6(1), 43-55.

Norris, C.M., Ghali, W.A., Knudtson, M.L., Naylor, C.D., and Saunders, L.D. (2000). Dealing with Missing Data in Observational Health Care Outcome Analyses. Journal of Clinical Epidemiology, 53(4), 377-378.

Nunamaker, T.R. (1983). Measuring Routine Nursing Service Efficiency: A Comparison of Cost per Patient Day and Data Envelopment Analysis Models. Health Services Research, 18 (2), 183-208.

Ozcan, Y.A. (1998). Physician Benchmarking: Measuring Variation in Practice Behavior in Treatment of Otitis Media. Health Care Management Science, 1(1), 5-17.

Pantall, J. (2001). Benchmarking in Healthcare. Nursing Times Research, 6(2), 568-580.

Paradi, J.C., Vela, S., and Yang, Z. (2004). Chapter 13: Assessing Bank and Bank Branch Performance. In Cooper, W.W., Seiford, L.M., and Zhu, J., Handbook on Data Envelopment Analysis (pp. 349-400). Boston: Kluwer Academic Publishers.

Prokopenko, J. (1987). Productivity Management: A Practical Handbook. Switzerland: International Labor Office.

Puenpatom, R.A., and Rosenman R. (2008). Efficiency of Thai Provincial Public Hospitals during the Introduction of Universal Health Coverage using Capitation. Health Care Management Science, 11(4), 319-338.

Ray, P.S., Aiyappan, H., Elam, M.E., Merritt, T.W. (2005). Application of Cluster Analysis in Marketing Management. International Journal of Industrial Engineering: Theory Applications and Practice, 12(2), 127-133.

Reducing Waste and Inefficiency in Health Care through Lean Process Redesign. (2009). Agency for Healthcare, 09(M028), 2011.

Roll, Y., Cook, W.D., and Golany, B. (1991). Controlling Factor Weights in Data Envelopment Analysis. IIE Transactions, 1(1), 2-9.

Roll, Y., and Golany, B. (1993). Alternate Methods of Treating Factor Weights in DEA. OMEGA-The International Journal of Management Science, 21(1), 99-109.

Sherman, H.D. (1984). Hospital Efficiency Measurement and Evaluation: Empirical Test of a New Technique. Medical Care, 22(10), 922-938.

Siddharthan, K., Ahern, M., and Rosenman, R. (2000). Data Envelopment Analysis to Determine Efficiencies of Health Maintenance Organizations. Health Care Management Science, 3(1), 23-29.

Siddiqui, S.A. (2005). Managerial Economics and Financial Analysis. New Delhi: New Age International.

Smirlis, Y.G., Maragos, E.K., and Despotis, D.K. (2006). Data Envelopment Analysis with Missing Values: An Interval DEA Approach. Applied Mathematics and Computation, 177 (1), 1-10.

Spath, H. (1980). Cluster Analysis Algorithms. United Kingdom: West Sussex, Ellis Horwood Limited.

Thanassoulis, E. (1993). A Comparison of Regression Analysis and Data Envelopment Analysis as Alternative Methods for Performance Assessments. The Journal of the Operational Research Society, 44(11), 1129-1144.

Thompson, R.G., Singleton, F.G., Thrall, R.M., and Smith, B.A. (1986). Comparative Site Evaluations for Locating a High-Energy Physics Lab in Texas. Interfaces, 16, 35-49.

Toloo, M. (2009). Short Communication on Classifying Inputs and Outputs in DEA: A revised model. European Journal of Operational Research, 198(1) 358-360.

Ward Jr., J. (1963). Hierarchical Grouping to Optimize an Objective Function. Journal of the American Statistical Association, 58(301), 236–244.

WHO Statistical Information System (WHOSIS). (2011). Retrieved September, 2011, from http://www.who.int/whosis/whostat/en/.

# Appendix A - Bibliography

Based on the Bibliography work done by Becker (2011) starting from the inception of DEA in 1978, the extracted list of healthcare related journals and the number of published research articles in the field of DEA are presented below:

| **Count** | **Journal Title** |
|---|---|
| 45 | Health Care Management Science |
| 17 | Health Policy |
| 17 | Health Services Research |
| 12 | Health Economics |
| 09 | European Journal of Health Economics |
| 09 | Health Services Management Research |
| 08 | Journal of Health Economics |
| 06 | Journal of Health Management |
| 06 | International Journal of Healthcare Technology and Management |
| 06 | Journal of Health and Human Resources Administration |
| 05 | Journal of Health Care Finance |
| 04 | Health Care Management Review |
| 03 | Hospital and Health Services Administration |
| 02 | The Health Care Supervisor |
| 02 | International Journal of Health Care Quality Assurance |
| 02 | International Journal of Health Care Finance and Economics |
| 02 | Journal of Mental Health Policy and Economics |
| 01 | The Journal of Behavioral Health Services and Research |
| 01 | The International Journal of Health Planning and Management |
| 01 | Research in Healthcare Financial Management |
| 01 | Journal of Health and Human Services Administration |
| 01 | Journal of Public Health Medicine |
| 01 | Journal of Public Health |

and many more………

There are many other researchers who made significant contributions updating the Bibliography for Data Envelopment Analysis. The following are the list of significant works:

Becker D. (2011). DEA Bib. Retrieved October, 2011, from http://www.deabib.org/journals.html#x6-160005.

Seiford, L.M. (1997). A bibliography for Data Envelopment Analysis (1978-1996). Annals of Operations Research, 73, 393-438.

Tavares, G. (2002). A bibliography of data envelopment analysis (1978-2001). RRR 01-02, RUTCOR, Rutgers Center for Operations Research, Rutgers University, Piscataway, New Jersey.

Emrouznejad, A., Parker, B.R., Tavares, G. (2008). Evaluation of research in efficiency and productivity: A survey and analysis of the first 30 years of scholarly literature in DEA. Socio-Economic Planning Sciences, 42, 151-157.

# Appendix B - List of KAMU Clinics

The following is the list of current 42 member clinics under Kansas Association of Medically Underserved Clinics.

| No | Name of the Clinic | Address | County |
|---|---|---|---|
| 1 | Atchison Community Health Clinic | 217 M Street, Atchison, KS, 66002 | Atchison |
| 2 | Center for Health and Wellness | 2707 E 21st Street, Wichita, KS, 67214 | Sedgwick |
| 3 | Cheyenne County Clinic, St. Francis | 221 West 1st St. Francis, KS, 67756 | Cheyenne |
| 4 | Children's Mercy West<br>The Cordell Meeks Jr. Clinic | 4313 State Avenue Kansas City, KS, 66102 | Wyandotte |
| 5 | Community Health Center of Southeast Kansas | 3011 N Michigan, Pittsburg, KS, 66762 | Crawford |
| 6 | Community Health Council of Wyandotte County | 755 Minnesota Avenue, 1st Floor<br>Kansas City, KS, 66101 | Wyandotte |
| 7 | Community Health Ministry Clinic | 903 6th Street, Wamego, KS, 66547 | Pottawatomie |
| 8 | Douglas County Dental Clinic | 316 Maine Street, Lawrence, KS, 66044 | Douglas |
| 9 | Duchesne Clinic | 636 Tauromee. Kansas City, KS, 66101 | Wyandotte |
| 10 | E.C. Tyree Health & Dental Clinic | 1525 N Lorraine, Wichita, KS, 67214 | Sedgwick |
| 11 | First Care Clinic, Inc. | 105 W 13th, Hays, KS, 67601 | Ellis |
| 12 | Flint Hills Community Clinic | 401 Houston, Manhattan, KS, 66502 | Riley |
| 13 | Flint Hills Community Health Center | 420 W 15th Street Emporia, KS, 66801 | Lyon |
| 14 | Grace Med Health Clinic, Inc. | 1122 N Topeka Wichita, KS, 67214 | Sedgwick |
| 15 | Greeley County Health Services | 504 E 6th Street Sharon Springs, KS, 67758 | Wallace |
| 16 | Guadalupe Clinic, Inc. | 940 S St. Francis Wichita, KS, 67211 | Sedgwick |
| 17 | Health Care Access, Inc. | 330 Maine Lawrence, KS, 66044 | Douglas |
| 18 | Health Ministries Clinic | 209 S Pine Street Newton, KS, 67114 | Harvey |
| 19 | Health Partnership Clinic | 7171 W 95th Street, Ste. 100<br>Overland Park, KS, 66212 | Johnson |
| 20 | Heart of KS Family Health Care | 1905 19th Street Great Bend, KS, 67530 | Barton |
| 21 | Heartland Medical Clinic | 1 Riverfront Plaza, #100 Lawrence, KS, 66044 | Douglas |

| | | | |
|---|---|---|---|
| 22 | Hunter Health Clinic | 2318 E Central Wichita, KS, 67214 | Sedgwick |
| 23 | Johnson County Health Department | 11875 S Sunset, Suite 300, Olathe, KS, 66061 | Johnson |
| 24 | KS Statewide Farm worker Health Program | 1000 SW Jackson, Ste. 340, Topeka, KS, 66612 | Shawnee |
| 25 | Konza Prairie Community Health Center | 361 Grant Avenue Junction City, KS, 66441 | Geary |
| 26 | KU Health Partners/Silver City Health Center | 1428 S 32nd, Ste. 100 Kansas City, KS, 66106 | Wyandotte |
| 27 | Marian Clinic | 1001 SW Garfield Avenue, Topeka, KS, 66604 | Shawnee |
| 28 | Mercy and Truth Medical Missions | 636 Minnesota Avenue Kansas City, KS, 66101 | Wyandotte |
| 29 | Mercy Health Systems: Arma, Cherryvale, and Linn County | 216 E 4th Cherryvale, KS, 67335 | Montgomery |
| 30 | Montgomery County Community Clinic (MC3) | 900 W Myrtle Independence. KS, 67301 | Montgomery |
| 31 | Mother Mary Anne Clinic | 1152 S Clifton Wichita, KS, 67218 | Sedgwick |
| 32 | Prairie Star Health Center | 1600 N Lorraine, St. 110 Hutchinson, KS, 67501 | Reno |
| 33 | Rawlins County Dental Clinic | 707 Grant Street, Atwood, KS, 67730 | Rawlins |
| 34 | Riley County Community Health Clinic | 2030 Tecumseh Manhattan, KS, 66502 | Riley |
| 35 | Saint Vincent Clinic | 818 N 7th Street Leavenworth, KS, 66048 | Leavenworth |
| 36 | Salina Family Health Care Center | 651 E Prescott Salina, KS, 67401 | Saline |
| 37 | Shawnee County Health Agency/CHC | 1615 SW 8th Street Topeka, KS, 66606 | Shawnee |
| 38 | Southwest Boulevard Family Health Care | 340 Southwest Boulevard Kansas City, KS, 66103 | Wyandotte |
| 39 | St. Gianna Health Clinic | 638 West D Avenue Kingman, KS, 67068 | Kingman |
| 40 | Swope Health Wyandotte and Swope Health West | 21 N 12th Street, Ste. 475 Kansas City, KS, 66102 | Wyandotte |
| 41 | Turner House Children's Clinic | 21 N 12th, Ste. 300 Kansas City, KS, 66102 | Wyandotte |
| 42 | United Methodist Mexican American Ministry | 712A St. John Street Garden City, KS, 67846 | Finney |

# Appendix C - Cluster Dispersion Example

Cluster Dispersion proposed by Himmelspach and Conrad (2010) tries to reduce the likelihood of marginal objects of a large cluster being falsely assigned to the nearest small cluster. Cluster dispersion values helps in updating the membership function. To understand how cluster dispersion updates the membership function in case of smaller and larger clusters and to understand its influence, let's consider a simple example of 8 observations classified into 2 clusters. This will be presented using two different kinds of datasets, one with fractional numbers and other with whole numbers.

## C.1 Fractional Numbers

Let's assume that out of 8 observations, cluster center 1 $(C_1)$ has 3 observations at a distance of **0.1** from the cluster center. Let's assume that the other 5 observations are also at a distance of **0.1** from cluster center 2, $(C_2)$. The can be represented using Figure C.1.

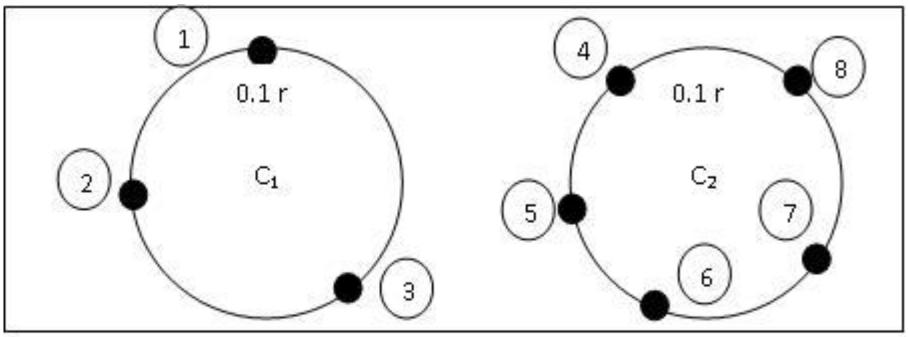

**Figure C-1: Clusters 1 and 2 (Fractional Numbers)**

According to Himmelspach and Conrad, cluster dispersion value, $S_k$, is calculated as follows:

$$S_k = \frac{1}{|v_k \cap X_{obs}| - 1} \sum_{x_j \in v_k \cap x_{obs}} \sum_{f \in f_{obs}} (x_i.f - v_k.f)^2$$

Let’s calculate the cluster dispersion using the above formulae for both these cluster represented by $S_1$ and $S_2$ for clusters 1 and 2 respectively.

$$S_1 = \frac{1}{(3-1)}[0.1 + 0.1 + 0.1] = \frac{0.3}{2} = 0.15$$

$$S_2 = \frac{1}{(5-1)}[0.1 + 0.1 + 0.1 + 0.1 + 0.1] = \frac{0.5}{4} = 0.125$$

From these cluster dispersion values we can infer that under similar conditions, cluster with more number of observations possess lesser cluster dispersion value.

According to Himmelspach and Conrad, New membership is updated as follows:

$$u_{ik}{}^{*} = \frac{\left( S_k \, D_{ik}{}^{1/(1-m)}\right)}{\left(\sum_{k=1}^{c}\left( S_k \, D_{ik}{}^{\frac{1}{1-m}}\right)\right)}$$

Now let’s determine how these cluster dispersion values influence the membership of the observations within clusters. Where $u_{11}$ represent the membership of observation (1) towards cluster 1, $u_{12}$ represent the membership of observation (1) towards cluster 2. Where $u_{41}$ represent the membership of observation (4) towards cluster 1, $u_{42}$ represent the membership of observation (4) towards cluster 2. Where $D_{11}$ represent the distance of observation (1) towards cluster 1, $D_{12}$ represent the distance of observation (1) towards cluster 2.

Let’s assume that observation (1) is 0.8 units away from cluster center 2 and similarly observation (4) is 0.8 units away from cluster center 1.

$$u_{11} = \frac{S_1/D_{11}}{[S_1/D_{11} + S_2/D_{12}]} = \frac{0.15/0.1}{[0.15/0.1 + 0.125/0.8]} = 0.9$$

$$u_{12} = \frac{S_2/D_{12}}{[S_1/D_{11} + S_2/D_{12}]} = \frac{0.125/0.8}{[0.15/0.1 + 0.125/0.8]} = 0.1$$

Let's determine how these cluster dispersion values influence the membership of the observations within cluster 2.

$$u_{41} = \frac{S_1/D_{41}}{[S_1/D_{41} + S_2/D_{42}]} = \frac{0.15/0.8}{[0.15/0.8 + 0.125/0.1]} = 0.13$$

$$u_{42} = \frac{S_2/D_{42}}{[S_1/D_{41} + S_2/D_{42}]} = \frac{0.125/0.1}{[0.15/0.8 + 0.125/0.1]} = 0.87$$

Based on the updated membership values in both cluster 1 and 2, the updated membership value of a cluster group with more observations is less when compared to other cluster group with less number of clusters.

## C.2 Whole Numbers

In the previous case we have seen how to determine the cluster dispersion values and update the membership matrix when the given data is fractional numbers. In this section we will focus on the similar topics but based on whole numbers.

Let's consider the same example of 8 observations classified into 2 clusters. Let's assume that cluster center 1 $(C_1)$ has 3 observations at a distance of **2** from the cluster center. Let's assume that the other 5 observations are also at a distance of **2** but they belong to cluster center 2, $(C_2)$. The can be represented using Figure C.2.

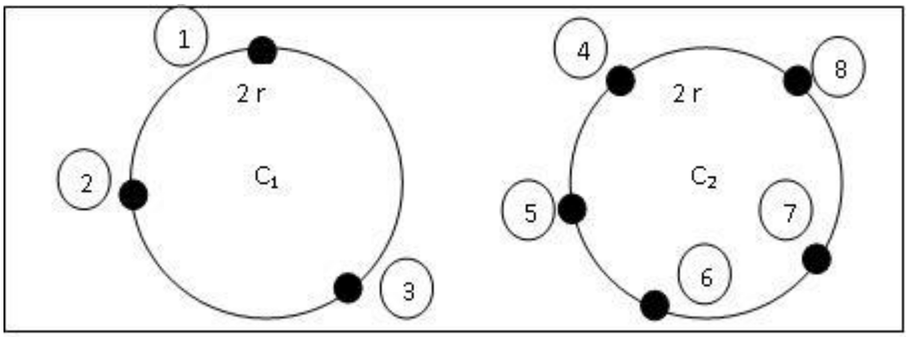

**Figure C-2: Clusters 1 and 2 (Whole Numbers)**

Let's calculate the cluster dispersion for both these cluster represented by $S_1$ and $S_2$ for cluster 1 and 2 respectively.

$$S_1 = \frac{1}{(3-1)}[2+2+2] = \frac{6}{2} = 3$$

$$S_2 = \frac{1}{(5-1)}[2+2+2+2+2] = \frac{10}{4} = 2.5$$

From these cluster dispersion values we can infer that under similar conditions, cluster with more number of observations possess lesser cluster dispersion.

Let's determine how these cluster dispersion values influence the membership of the observations within clusters.

Let's assume that observation (1) is 5 units away from cluster center 2 and similarly observation (4) is 5 units away from cluster center 1.

$$u_{11} = \frac{S_1/D_{11}}{[S_1/D_{11} + S_2/D_{12}]} = \frac{3/2}{[3/2 + 2.5/5]} = 0.5 * 1.5 = 0.75$$

$$u_{12} = \frac{S_2/D_{12}}{[S_1/D_{11} + S_2/D_{12}]} = \frac{2.5/5}{[3/2 + 2.5/5]} = 0.5 * 0.5 = 0.25$$

$$u_{41} = \frac{S_1/D_{41}}{[S_1/D_{41} + S_2/D_{42}]} = \frac{3/5}{[3/5 + 2.5/2]} = 0.54 * 0.6 = 0.324$$

$$u_{42} = \frac{S_2/D_{42}}{[S_1/D_{41} + S_2/D_{42}]} = \frac{2.5/2}{[3/5 + 2.5/2]} = 0.54 * 1.25 = 0.676$$

Based on the updated membership values in both cluster 1 and 2, the updated membership value of a cluster group with more observations is less when compared to other cluster group with less number of clusters.

# Appendix D - Relative and Fixed Intervals

This appendix presents the values assumed as missing in KAMU complete dataset of 22 clinics with 7 variables for each case 10%, 20% and 30% missing values. The relative and fixed intervals are constructed around the center value (actual known value). These intervals are broken down into crisp values based on linear interpolations using the concept of common value of alpha (α). Table D.1 to D.6 presents the crisp values that are used to replace the assumed missing values. Alpha of 0 represents the lower bound and alpha of 1 represents the upper bound of the intervals. Table D.1 to D.3 represent the relative intervals for 10%, 20% and 30% missing values cases respectively. Table D.4 to D.6 represent the fixed intervals.

**Table D-1: Relative Intervals in case of 10% Missing Values**

| **Reference Cells** | **Actual Values** | **Alpha Values** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| X(1,6) | **0.040** | 0.035 | 0.036 | 0.037 | 0.038 | 0.039 | 0.040 | 0.041 | 0.042 | 0.043 | 0.044 | 0.045 |
| X(2,2) | **0.461** | 0.403 | 0.415 | 0.426 | 0.438 | 0.449 | 0.461 | 0.472 | 0.484 | 0.495 | 0.507 | 0.518 |
| X(3,7) | **0.170** | 0.149 | 0.153 | 0.157 | 0.162 | 0.166 | 0.170 | 0.174 | 0.179 | 0.183 | 0.187 | 0.191 |
| X(5,3) | **0.574** | 0.502 | 0.516 | 0.530 | 0.545 | 0.559 | 0.574 | 0.588 | 0.602 | 0.617 | 0.631 | 0.645 |
| X(6,5) | **0.155** | 0.135 | 0.139 | 0.143 | 0.147 | 0.151 | 0.155 | 0.159 | 0.162 | 0.166 | 0.170 | 0.174 |
| X(8,1) | **0.115** | 0.101 | 0.104 | 0.106 | 0.109 | 0.112 | 0.115 | 0.118 | 0.121 | 0.124 | 0.127 | 0.129 |
| X(9,7) | **0.118** | 0.103 | 0.106 | 0.109 | 0.112 | 0.115 | 0.118 | 0.121 | 0.124 | 0.127 | 0.130 | 0.132 |
| X(12,2) | **1.000** | 0.750 | 0.775 | 0.800 | 0.825 | 0.850 | 0.875 | 0.900 | 0.925 | 0.950 | 0.975 | 1.000 |
| X(13,7) | **0.319** | 0.279 | 0.287 | 0.295 | 0.303 | 0.311 | 0.319 | 0.327 | 0.335 | 0.343 | 0.351 | 0.359 |
| X(15,4) | **0.201** | 0.176 | 0.181 | 0.186 | 0.191 | 0.196 | 0.201 | 0.206 | 0.211 | 0.216 | 0.221 | 0.226 |
| X(16,5) | **0.602** | 0.527 | 0.542 | 0.557 | 0.572 | 0.587 | 0.602 | 0.617 | 0.632 | 0.647 | 0.662 | 0.677 |
| X(18,3) | **0.055** | 0.048 | 0.050 | 0.051 | 0.052 | 0.054 | 0.055 | 0.057 | 0.058 | 0.059 | 0.061 | 0.062 |
| X(19,4) | **0.066** | 0.058 | 0.060 | 0.061 | 0.063 | 0.065 | 0.066 | 0.068 | 0.070 | 0.071 | 0.073 | 0.075 |
| X(20,6) | **0.609** | 0.532 | 0.548 | 0.563 | 0.578 | 0.593 | 0.609 | 0.624 | 0.639 | 0.654 | 0.669 | 0.685 |
| X(21,1) | **0.450** | 0.394 | 0.405 | 0.417 | 0.428 | 0.439 | 0.450 | 0.462 | 0.473 | 0.484 | 0.496 | 0.507 |
| X(22,7) | **0.263** | 0.230 | 0.236 | 0.243 | 0.250 | 0.256 | 0.263 | 0.269 | 0.276 | 0.282 | 0.289 | 0.296 |
| **Mean Abs Dev** | | **0.048** | **0.040** | **0.032** | **0.024** | **0.016** | **0.008** | **0.013** | **0.018** | **0.023** | **0.028** | **0.033** |

### Table D-2: Relative Intervals in case of 20% Missing Values

| Reference Cells | Actual Values | Alpha Values | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| X(1,3) | 0.138 | 0.104 | 0.110 | 0.117 | 0.124 | 0.131 | 0.138 | 0.145 | 0.152 | 0.159 | 0.166 | 0.173 |
| X(1,4) | 0.066 | 0.050 | 0.053 | 0.056 | 0.060 | 0.063 | 0.066 | 0.070 | 0.073 | 0.076 | 0.080 | 0.083 |
| X(2,5) | 0.458 | 0.343 | 0.366 | 0.389 | 0.412 | 0.435 | 0.458 | 0.480 | 0.503 | 0.526 | 0.549 | 0.572 |
| X(3,1) | 0.054 | 0.040 | 0.043 | 0.046 | 0.048 | 0.051 | 0.054 | 0.056 | 0.059 | 0.062 | 0.064 | 0.067 |
| X(4,6) | 0.060 | 0.045 | 0.048 | 0.051 | 0.054 | 0.057 | 0.060 | 0.063 | 0.066 | 0.069 | 0.072 | 0.075 |
| X(5,4) | 1.000 | 0.500 | 0.550 | 0.600 | 0.650 | 0.700 | 0.750 | 0.800 | 0.850 | 0.900 | 0.950 | 1.000 |
| X(5,5) | 0.474 | 0.356 | 0.379 | 0.403 | 0.427 | 0.450 | 0.474 | 0.498 | 0.521 | 0.545 | 0.569 | 0.593 |
| X(6,6) | 0.128 | 0.096 | 0.102 | 0.109 | 0.115 | 0.121 | 0.128 | 0.134 | 0.141 | 0.147 | 0.153 | 0.160 |
| X(6,7) | 0.154 | 0.115 | 0.123 | 0.131 | 0.138 | 0.146 | 0.154 | 0.161 | 0.169 | 0.177 | 0.184 | 0.192 |
| X(7,1) | 0.492 | 0.369 | 0.394 | 0.419 | 0.443 | 0.468 | 0.492 | 0.517 | 0.542 | 0.566 | 0.591 | 0.615 |
| X(7,3) | 0.615 | 0.461 | 0.492 | 0.523 | 0.553 | 0.584 | 0.615 | 0.646 | 0.676 | 0.707 | 0.738 | 0.769 |
| X(8,2) | 0.530 | 0.398 | 0.424 | 0.451 | 0.477 | 0.504 | 0.530 | 0.557 | 0.583 | 0.610 | 0.636 | 0.663 |
| X(8,5) | 0.170 | 0.128 | 0.136 | 0.145 | 0.153 | 0.162 | 0.170 | 0.179 | 0.187 | 0.196 | 0.204 | 0.213 |
| X(9,7) | 0.118 | 0.088 | 0.094 | 0.100 | 0.106 | 0.112 | 0.118 | 0.124 | 0.130 | 0.135 | 0.141 | 0.147 |
| X(10,5) | 0.114 | 0.086 | 0.092 | 0.097 | 0.103 | 0.109 | 0.114 | 0.120 | 0.126 | 0.132 | 0.137 | 0.143 |
| X(11,5) | 0.094 | 0.070 | 0.075 | 0.080 | 0.084 | 0.089 | 0.094 | 0.098 | 0.103 | 0.108 | 0.112 | 0.117 |
| X(12,4) | 0.716 | 0.537 | 0.573 | 0.609 | 0.645 | 0.680 | 0.716 | 0.752 | 0.788 | 0.824 | 0.860 | 0.895 |
| X(12,7) | 0.387 | 0.290 | 0.310 | 0.329 | 0.348 | 0.368 | 0.387 | 0.406 | 0.426 | 0.445 | 0.464 | 0.484 |
| X(13,2) | 0.266 | 0.199 | 0.213 | 0.226 | 0.239 | 0.253 | 0.266 | 0.279 | 0.292 | 0.306 | 0.319 | 0.332 |
| X(14,1) | 0.264 | 0.198 | 0.211 | 0.224 | 0.237 | 0.251 | 0.264 | 0.277 | 0.290 | 0.303 | 0.317 | 0.330 |
| X(14,6) | 0.033 | 0.024 | 0.026 | 0.028 | 0.029 | 0.031 | 0.033 | 0.034 | 0.036 | 0.037 | 0.039 | 0.041 |
| X(15,5) | 0.268 | 0.201 | 0.215 | 0.228 | 0.242 | 0.255 | 0.268 | 0.282 | 0.295 | 0.309 | 0.322 | 0.335 |
| X(16,1) | 0.411 | 0.308 | 0.329 | 0.349 | 0.370 | 0.390 | 0.411 | 0.431 | 0.452 | 0.472 | 0.493 | 0.513 |
| X(16,7) | 0.158 | 0.119 | 0.126 | 0.134 | 0.142 | 0.150 | 0.158 | 0.166 | 0.174 | 0.182 | 0.190 | 0.198 |
| X(17,6) | 0.378 | 0.283 | 0.302 | 0.321 | 0.340 | 0.359 | 0.378 | 0.397 | 0.416 | 0.435 | 0.453 | 0.472 |
| X(18,3) | 0.055 | 0.041 | 0.044 | 0.047 | 0.050 | 0.052 | 0.055 | 0.058 | 0.061 | 0.063 | 0.066 | 0.069 |
| X(19,2) | 0.162 | 0.122 | 0.130 | 0.138 | 0.146 | 0.154 | 0.162 | 0.170 | 0.178 | 0.186 | 0.194 | 0.203 |
| X(19,6) | 0.132 | 0.099 | 0.106 | 0.112 | 0.119 | 0.125 | 0.132 | 0.139 | 0.145 | 0.152 | 0.158 | 0.165 |
| X(20,4) | 0.133 | 0.100 | 0.106 | 0.113 | 0.120 | 0.126 | 0.133 | 0.140 | 0.146 | 0.153 | 0.159 | 0.166 |
| X(21,2) | 0.193 | 0.145 | 0.154 | 0.164 | 0.174 | 0.183 | 0.193 | 0.203 | 0.212 | 0.222 | 0.232 | 0.241 |
| X(22,3) | 0.328 | 0.246 | 0.262 | 0.279 | 0.295 | 0.311 | 0.328 | 0.344 | 0.361 | 0.377 | 0.393 | 0.410 |
| X(22,7) | 0.263 | 0.197 | 0.210 | 0.223 | 0.236 | 0.250 | 0.263 | 0.276 | 0.289 | 0.302 | 0.315 | 0.328 |
| **Mean Abs Dev** | | **0.077** | **0.063** | **0.049** | **0.035** | **0.022** | **0.008** | **0.018** | **0.029** | **0.040** | **0.050** | **0.061** |

### Table D-3: Relative Intervals in case of 30% Missing Values

| Reference Cells | Actual Values | Alpha Values | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1 |
| X(1,1) | 0.127 | 0.095 | 0.102 | 0.108 | 0.115 | 0.121 | 0.127 | 0.134 | 0.140 | 0.146 | 0.153 | 0.159 |
| X(1,6) | 0.040 | 0.030 | 0.032 | 0.034 | 0.036 | 0.038 | 0.040 | 0.042 | 0.044 | 0.046 | 0.048 | 0.050 |
| X(2,3) | 0.766 | 0.575 | 0.613 | 0.651 | 0.689 | 0.728 | 0.766 | 0.804 | 0.843 | 0.881 | 0.919 | 0.958 |
| X(2,5) | 0.458 | 0.343 | 0.366 | 0.389 | 0.412 | 0.435 | 0.458 | 0.480 | 0.503 | 0.526 | 0.549 | 0.572 |
| X(3,1) | 0.054 | 0.040 | 0.043 | 0.046 | 0.048 | 0.051 | 0.054 | 0.056 | 0.059 | 0.062 | 0.064 | 0.067 |
| X(3,7) | 0.170 | 0.128 | 0.136 | 0.145 | 0.153 | 0.162 | 0.170 | 0.179 | 0.187 | 0.196 | 0.204 | 0.213 |
| X(4,3) | 0.207 | 0.155 | 0.166 | 0.176 | 0.186 | 0.197 | 0.207 | 0.217 | 0.228 | 0.238 | 0.248 | 0.259 |
| X(4,4) | 0.066 | 0.050 | 0.053 | 0.056 | 0.060 | 0.063 | 0.066 | 0.070 | 0.073 | 0.076 | 0.080 | 0.083 |
| X(4,7) | 0.187 | 0.141 | 0.150 | 0.159 | 0.169 | 0.178 | 0.187 | 0.197 | 0.206 | 0.215 | 0.225 | 0.234 |
| X(5,2) | 0.444 | 0.333 | 0.355 | 0.377 | 0.399 | 0.421 | 0.444 | 0.466 | 0.488 | 0.510 | 0.532 | 0.554 |
| X(5,5) | 0.474 | 0.356 | 0.379 | 0.403 | 0.427 | 0.450 | 0.474 | 0.498 | 0.521 | 0.545 | 0.569 | 0.593 |
| X(6,1) | 0.061 | 0.046 | 0.049 | 0.052 | 0.055 | 0.058 | 0.061 | 0.064 | 0.067 | 0.070 | 0.073 | 0.076 |
| X(6,5) | 0.155 | 0.116 | 0.124 | 0.132 | 0.139 | 0.147 | 0.155 | 0.162 | 0.170 | 0.178 | 0.186 | 0.193 |
| X(6,6) | 0.128 | 0.096 | 0.102 | 0.109 | 0.115 | 0.121 | 0.128 | 0.134 | 0.141 | 0.147 | 0.153 | 0.160 |
| X(7,2) | 0.690 | 0.517 | 0.552 | 0.586 | 0.621 | 0.655 | 0.690 | 0.724 | 0.759 | 0.793 | 0.828 | 0.862 |
| X(7,4) | 0.160 | 0.120 | 0.128 | 0.136 | 0.144 | 0.152 | 0.160 | 0.168 | 0.176 | 0.184 | 0.192 | 0.200 |
| X(7,7) | 0.344 | 0.258 | 0.275 | 0.292 | 0.309 | 0.326 | 0.344 | 0.361 | 0.378 | 0.395 | 0.412 | 0.430 |
| X(8,3) | 0.138 | 0.104 | 0.110 | 0.117 | 0.124 | 0.131 | 0.138 | 0.145 | 0.152 | 0.159 | 0.166 | 0.173 |
| X(8,6) | 0.014 | 0.011 | 0.011 | 0.012 | 0.013 | 0.014 | 0.014 | 0.015 | 0.016 | 0.016 | 0.017 | 0.018 |
| X(9,1) | 0.070 | 0.053 | 0.056 | 0.060 | 0.063 | 0.067 | 0.070 | 0.074 | 0.078 | 0.081 | 0.085 | 0.088 |
| X(9,7) | 0.118 | 0.088 | 0.094 | 0.100 | 0.106 | 0.112 | 0.118 | 0.124 | 0.130 | 0.135 | 0.141 | 0.147 |
| X(10,4) | 0.159 | 0.120 | 0.128 | 0.136 | 0.144 | 0.151 | 0.159 | 0.167 | 0.175 | 0.183 | 0.191 | 0.199 |
| X(10,5) | 0.114 | 0.086 | 0.092 | 0.097 | 0.103 | 0.109 | 0.114 | 0.120 | 0.126 | 0.132 | 0.137 | 0.143 |
| X(11,2) | 0.198 | 0.149 | 0.159 | 0.169 | 0.179 | 0.189 | 0.198 | 0.208 | 0.218 | 0.228 | 0.238 | 0.248 |
| X(11,6) | 0.039 | 0.029 | 0.031 | 0.033 | 0.035 | 0.037 | 0.039 | 0.041 | 0.043 | 0.045 | 0.047 | 0.049 |
| X(12,4) | 0.716 | 0.537 | 0.573 | 0.609 | 0.645 | 0.680 | 0.716 | 0.752 | 0.788 | 0.824 | 0.860 | 0.895 |
| X(12,7) | 0.387 | 0.290 | 0.310 | 0.329 | 0.348 | 0.368 | 0.387 | 0.406 | 0.426 | 0.445 | 0.464 | 0.484 |
| X(13,2) | 0.266 | 0.199 | 0.213 | 0.226 | 0.239 | 0.253 | 0.266 | 0.279 | 0.292 | 0.306 | 0.319 | 0.332 |
| X(13,5) | 0.175 | 0.131 | 0.140 | 0.149 | 0.158 | 0.166 | 0.175 | 0.184 | 0.193 | 0.201 | 0.210 | 0.219 |
| X(14,1) | 0.264 | 0.198 | 0.211 | 0.224 | 0.237 | 0.251 | 0.264 | 0.277 | 0.290 | 0.303 | 0.317 | 0.330 |
| X(14,3) | 0.255 | 0.192 | 0.204 | 0.217 | 0.230 | 0.243 | 0.255 | 0.268 | 0.281 | 0.294 | 0.306 | 0.319 |
| X(15,4) | 0.201 | 0.151 | 0.161 | 0.171 | 0.181 | 0.191 | 0.201 | 0.211 | 0.221 | 0.232 | 0.242 | 0.252 |
| X(15,7) | 0.217 | 0.162 | 0.173 | 0.184 | 0.195 | 0.206 | 0.217 | 0.227 | 0.238 | 0.249 | 0.260 | 0.271 |
| X(16,3) | 0.407 | 0.305 | 0.326 | 0.346 | 0.366 | 0.387 | 0.407 | 0.428 | 0.448 | 0.468 | 0.489 | 0.509 |
| X(16,6) | 0.287 | 0.215 | 0.229 | 0.244 | 0.258 | 0.272 | 0.287 | 0.301 | 0.315 | 0.330 | 0.344 | 0.358 |
| X(17,1) | 0.683 | 0.512 | 0.546 | 0.580 | 0.614 | 0.649 | 0.683 | 0.717 | 0.751 | 0.785 | 0.819 | 0.853 |
| X(17,5) | 0.421 | 0.316 | 0.337 | 0.358 | 0.379 | 0.400 | 0.421 | 0.443 | 0.464 | 0.485 | 0.506 | 0.527 |
| X(18,2) | 0.215 | 0.161 | 0.172 | 0.183 | 0.193 | 0.204 | 0.215 | 0.226 | 0.236 | 0.247 | 0.258 | 0.268 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X(18,7) | 0.176 | 0.132 | 0.140 | 0.149 | 0.158 | 0.167 | 0.176 | 0.184 | 0.193 | 0.202 | 0.211 | 0.219 |
| X(19,4) | 0.066 | 0.050 | 0.053 | 0.056 | 0.060 | 0.063 | 0.066 | 0.070 | 0.073 | 0.076 | 0.080 | 0.083 |
| X(19,5) | 0.150 | 0.112 | 0.120 | 0.127 | 0.135 | 0.142 | 0.150 | 0.157 | 0.165 | 0.172 | 0.180 | 0.187 |
| X(19,6) | 0.132 | 0.099 | 0.106 | 0.112 | 0.119 | 0.125 | 0.132 | 0.139 | 0.145 | 0.152 | 0.158 | 0.165 |
| X(20,2) | 0.323 | 0.243 | 0.259 | 0.275 | 0.291 | 0.307 | 0.323 | 0.340 | 0.356 | 0.372 | 0.388 | 0.404 |
| X(20,6) | 0.609 | 0.456 | 0.487 | 0.517 | 0.548 | 0.578 | 0.609 | 0.639 | 0.669 | 0.700 | 0.730 | 0.761 |
| X(21,7) | 0.176 | 0.132 | 0.140 | 0.149 | 0.158 | 0.167 | 0.176 | 0.184 | 0.193 | 0.202 | 0.211 | 0.219 |
| X(21,3) | 0.207 | 0.155 | 0.166 | 0.176 | 0.186 | 0.197 | 0.207 | 0.217 | 0.228 | 0.238 | 0.248 | 0.259 |
| X(22,5) | 0.195 | 0.146 | 0.156 | 0.166 | 0.176 | 0.185 | 0.195 | 0.205 | 0.215 | 0.225 | 0.234 | 0.244 |
| X(22,6) | 0.239 | 0.179 | 0.191 | 0.203 | 0.215 | 0.227 | 0.239 | 0.251 | 0.263 | 0.275 | 0.287 | 0.299 |
| **Mean Abs Dev** | | **0.063** | **0.051** | **0.038** | **0.025** | **0.013** | **0.000** | **0.013** | **0.025** | **0.038** | **0.051** | **0.063** |

## Table D-4: Fixed Intervals in case of 10% Missing Values

| Reference Cells | Actual Values | Alpha Values | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| X(1,6) | **0.040** | 0.000 | 0.029 | 0.058 | 0.087 | 0.116 | 0.145 | 0.174 | 0.203 | 0.232 | 0.261 | 0.290 |
| X(2,2) | **0.461** | 0.211 | 0.261 | 0.311 | 0.361 | 0.411 | 0.461 | 0.511 | 0.561 | 0.611 | 0.661 | 0.711 |
| X(3,7) | **0.170** | 0.000 | 0.042 | 0.084 | 0.126 | 0.168 | 0.210 | 0.252 | 0.294 | 0.336 | 0.378 | 0.420 |
| X(5,3) | **0.574** | 0.324 | 0.374 | 0.424 | 0.474 | 0.524 | 0.574 | 0.624 | 0.674 | 0.724 | 0.774 | 0.824 |
| X(6,5) | **0.155** | 0.000 | 0.040 | 0.081 | 0.121 | 0.162 | 0.202 | 0.243 | 0.283 | 0.324 | 0.364 | 0.405 |
| X(8,1) | **0.115** | 0.000 | 0.037 | 0.073 | 0.110 | 0.146 | 0.183 | 0.219 | 0.256 | 0.292 | 0.329 | 0.365 |
| X(9,7) | **0.118** | 0.000 | 0.037 | 0.074 | 0.110 | 0.147 | 0.184 | 0.221 | 0.257 | 0.294 | 0.331 | 0.368 |
| X(12,2) | **1.000** | 0.750 | 0.775 | 0.800 | 0.825 | 0.850 | 0.875 | 0.900 | 0.925 | 0.950 | 0.975 | 1.000 |
| X(13,7) | **0.319** | 0.069 | 0.119 | 0.169 | 0.219 | 0.269 | 0.319 | 0.369 | 0.419 | 0.469 | 0.519 | 0.569 |
| X(15,4) | **0.201** | 0.000 | 0.045 | 0.090 | 0.135 | 0.181 | 0.226 | 0.271 | 0.316 | 0.361 | 0.406 | 0.451 |
| X(16,5) | **0.602** | 0.352 | 0.402 | 0.452 | 0.502 | 0.552 | 0.602 | 0.652 | 0.702 | 0.752 | 0.802 | 0.852 |
| X(18,3) | **0.055** | 0.000 | 0.031 | 0.061 | 0.092 | 0.122 | 0.153 | 0.183 | 0.214 | 0.244 | 0.275 | 0.305 |
| X(19,4) | **0.066** | 0.000 | 0.032 | 0.063 | 0.095 | 0.127 | 0.158 | 0.190 | 0.222 | 0.253 | 0.285 | 0.316 |
| X(20,6) | **0.609** | 0.359 | 0.409 | 0.459 | 0.509 | 0.559 | 0.609 | 0.659 | 0.709 | 0.759 | 0.809 | 0.859 |
| X(21,1) | **0.450** | 0.200 | 0.250 | 0.300 | 0.350 | 0.400 | 0.450 | 0.500 | 0.550 | 0.600 | 0.650 | 0.700 |
| X(22,7) | **0.263** | 0.013 | 0.063 | 0.113 | 0.163 | 0.213 | 0.263 | 0.313 | 0.363 | 0.413 | 0.463 | 0.513 |
| **Mean Abs Dev** | | **0.183** | **0.141** | **0.102** | **0.071** | **0.050** | **0.042** | **0.080** | **0.119** | **0.157** | **0.196** | **0.234** |

## Table D-5: Fixed Intervals in case of 20% Missing Values

| Reference Cells | Actual Values | Alpha Values | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| X(1,3) | 0.138 | 0.000 | 0.039 | 0.078 | 0.116 | 0.155 | 0.194 | 0.233 | 0.272 | 0.310 | 0.349 | 0.388 |
| X(1,4) | 0.066 | 0.000 | 0.032 | 0.063 | 0.095 | 0.127 | 0.158 | 0.190 | 0.222 | 0.253 | 0.285 | 0.316 |
| X(2,5) | 0.458 | 0.208 | 0.258 | 0.308 | 0.358 | 0.408 | 0.458 | 0.508 | 0.558 | 0.608 | 0.658 | 0.708 |
| X(3,1) | 0.054 | 0.000 | 0.030 | 0.061 | 0.091 | 0.121 | 0.152 | 0.182 | 0.213 | 0.243 | 0.273 | 0.304 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X(4,6) | 0.060 | 0.000 | 0.031 | 0.062 | 0.093 | 0.124 | 0.155 | 0.186 | 0.217 | 0.248 | 0.279 | 0.310 |
| X(5,4) | 1.000 | 0.750 | 0.775 | 0.800 | 0.825 | 0.850 | 0.875 | 0.900 | 0.925 | 0.950 | 0.975 | 1.000 |
| X(5,5) | 0.474 | 0.224 | 0.274 | 0.324 | 0.374 | 0.424 | 0.474 | 0.524 | 0.574 | 0.624 | 0.674 | 0.724 |
| X(6,6) | 0.128 | 0.000 | 0.038 | 0.076 | 0.113 | 0.151 | 0.189 | 0.227 | 0.264 | 0.302 | 0.340 | 0.378 |
| X(6,7) | 0.154 | 0.000 | 0.040 | 0.081 | 0.121 | 0.161 | 0.202 | 0.242 | 0.283 | 0.323 | 0.363 | 0.404 |
| X(7,1) | 0.492 | 0.242 | 0.292 | 0.342 | 0.392 | 0.442 | 0.492 | 0.542 | 0.592 | 0.642 | 0.692 | 0.742 |
| X(7,3) | 0.615 | 0.365 | 0.415 | 0.465 | 0.515 | 0.565 | 0.615 | 0.665 | 0.715 | 0.765 | 0.815 | 0.865 |
| X(8,2) | 0.530 | 0.280 | 0.330 | 0.380 | 0.430 | 0.480 | 0.530 | 0.580 | 0.630 | 0.680 | 0.730 | 0.780 |
| X(8,5) | 0.170 | 0.000 | 0.042 | 0.084 | 0.126 | 0.168 | 0.210 | 0.252 | 0.294 | 0.336 | 0.378 | 0.420 |
| X(9,7) | 0.118 | 0.000 | 0.037 | 0.074 | 0.110 | 0.147 | 0.184 | 0.221 | 0.257 | 0.294 | 0.331 | 0.368 |
| X(10,5) | 0.114 | 0.000 | 0.036 | 0.073 | 0.109 | 0.146 | 0.182 | 0.219 | 0.255 | 0.292 | 0.328 | 0.364 |
| X(11,5) | 0.094 | 0.000 | 0.034 | 0.069 | 0.103 | 0.137 | 0.172 | 0.206 | 0.241 | 0.275 | 0.309 | 0.344 |
| X(12,4) | 0.716 | 0.466 | 0.516 | 0.566 | 0.616 | 0.666 | 0.716 | 0.766 | 0.816 | 0.866 | 0.916 | 0.966 |
| X(12,7) | 0.387 | 0.137 | 0.187 | 0.237 | 0.287 | 0.337 | 0.387 | 0.437 | 0.487 | 0.537 | 0.587 | 0.637 |
| X(13,2) | 0.266 | 0.016 | 0.066 | 0.116 | 0.166 | 0.216 | 0.266 | 0.316 | 0.366 | 0.416 | 0.466 | 0.516 |
| X(14,1) | 0.264 | 0.014 | 0.064 | 0.114 | 0.164 | 0.214 | 0.264 | 0.314 | 0.364 | 0.414 | 0.464 | 0.514 |
| X(14,6) | 0.033 | 0.000 | 0.028 | 0.057 | 0.085 | 0.113 | 0.141 | 0.170 | 0.198 | 0.226 | 0.254 | 0.283 |
| X(15,5) | 0.268 | 0.018 | 0.068 | 0.118 | 0.168 | 0.218 | 0.268 | 0.318 | 0.368 | 0.418 | 0.468 | 0.518 |
| X(16,1) | 0.411 | 0.161 | 0.211 | 0.261 | 0.311 | 0.361 | 0.411 | 0.461 | 0.511 | 0.561 | 0.611 | 0.661 |
| X(16,7) | 0.158 | 0.000 | 0.041 | 0.082 | 0.122 | 0.163 | 0.204 | 0.245 | 0.286 | 0.326 | 0.367 | 0.408 |
| X(17,6) | 0.378 | 0.128 | 0.178 | 0.228 | 0.278 | 0.328 | 0.378 | 0.428 | 0.478 | 0.528 | 0.578 | 0.628 |
| X(18,3) | 0.055 | 0.000 | 0.031 | 0.061 | 0.092 | 0.122 | 0.153 | 0.183 | 0.214 | 0.244 | 0.275 | 0.305 |
| X(19,2) | 0.162 | 0.000 | 0.041 | 0.082 | 0.124 | 0.165 | 0.206 | 0.247 | 0.288 | 0.330 | 0.371 | 0.412 |
| X(19,6) | 0.132 | 0.000 | 0.038 | 0.076 | 0.115 | 0.153 | 0.191 | 0.229 | 0.267 | 0.306 | 0.344 | 0.382 |
| X(20,4) | 0.133 | 0.000 | 0.038 | 0.077 | 0.115 | 0.153 | 0.191 | 0.230 | 0.268 | 0.306 | 0.345 | 0.383 |
| X(21,2) | 0.193 | 0.000 | 0.044 | 0.089 | 0.133 | 0.177 | 0.222 | 0.266 | 0.310 | 0.354 | 0.399 | 0.443 |
| X(22,3) | 0.328 | 0.078 | 0.128 | 0.178 | 0.228 | 0.278 | 0.328 | 0.378 | 0.428 | 0.478 | 0.528 | 0.578 |
| X(22,7) | 0.263 | 0.013 | 0.063 | 0.113 | 0.163 | 0.213 | 0.263 | 0.313 | 0.363 | 0.413 | 0.463 | 0.513 |
| **Mean Abs Dev** | | **0.178** | **0.136** | **0.097** | **0.065** | **0.044** | **0.040** | **0.080** | **0.121** | **0.161** | **0.202** | **0.242** |

**Table D-6: Fixed Intervals in case of 30% Missing Values**

| **Reference Cells** | **Actual Values** | **Alpha Values** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **0** | **0.1** | **0.2** | **0.3** | **0.4** | **0.5** | **0.6** | **0.7** | **0.8** | **0.9** | **1** |
| X(1,1) | 0.127 | 0.000 | 0.038 | 0.075 | 0.113 | 0.151 | 0.189 | 0.226 | 0.264 | 0.302 | 0.340 | 0.377 |
| X(1,6) | 0.040 | 0.000 | 0.029 | 0.058 | 0.087 | 0.116 | 0.145 | 0.174 | 0.203 | 0.232 | 0.261 | 0.290 |
| X(2,3) | 0.766 | 0.516 | 0.564 | 0.613 | 0.661 | 0.710 | 0.758 | 0.806 | 0.855 | 0.903 | 0.952 | 1.000 |
| X(2,5) | 0.458 | 0.208 | 0.258 | 0.308 | 0.358 | 0.408 | 0.458 | 0.508 | 0.558 | 0.608 | 0.658 | 0.708 |
| X(3,1) | 0.054 | 0.000 | 0.030 | 0.061 | 0.091 | 0.121 | 0.152 | 0.182 | 0.213 | 0.243 | 0.273 | 0.304 |
| X(3,7) | 0.170 | 0.000 | 0.042 | 0.084 | 0.126 | 0.168 | 0.210 | 0.252 | 0.294 | 0.336 | 0.378 | 0.420 |
| X(4,3) | 0.207 | 0.000 | 0.046 | 0.091 | 0.137 | 0.183 | 0.229 | 0.274 | 0.320 | 0.366 | 0.411 | 0.457 |
| X(4,4) | 0.066 | 0.000 | 0.032 | 0.063 | 0.095 | 0.127 | 0.158 | 0.190 | 0.222 | 0.253 | 0.285 | 0.316 |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| X(4,7) | 0.187 | 0.000 | 0.044 | 0.087 | 0.131 | 0.175 | 0.219 | 0.262 | 0.306 | 0.350 | 0.394 | 0.437 |
| X(5,2) | 0.444 | 0.194 | 0.244 | 0.294 | 0.344 | 0.394 | 0.444 | 0.494 | 0.544 | 0.594 | 0.644 | 0.694 |
| X(5,5) | 0.474 | 0.224 | 0.274 | 0.324 | 0.374 | 0.424 | 0.474 | 0.524 | 0.574 | 0.624 | 0.674 | 0.724 |
| X(6,1) | 0.061 | 0.000 | 0.031 | 0.062 | 0.093 | 0.124 | 0.155 | 0.187 | 0.218 | 0.249 | 0.280 | 0.311 |
| X(6,5) | 0.155 | 0.000 | 0.040 | 0.081 | 0.121 | 0.162 | 0.202 | 0.243 | 0.283 | 0.324 | 0.364 | 0.405 |
| X(6,6) | 0.128 | 0.000 | 0.038 | 0.076 | 0.113 | 0.151 | 0.189 | 0.227 | 0.264 | 0.302 | 0.340 | 0.378 |
| X(7,2) | 0.690 | 0.440 | 0.490 | 0.540 | 0.590 | 0.640 | 0.690 | 0.740 | 0.790 | 0.840 | 0.890 | 0.940 |
| X(7,4) | 0.160 | 0.000 | 0.041 | 0.082 | 0.123 | 0.164 | 0.205 | 0.246 | 0.287 | 0.328 | 0.369 | 0.410 |
| X(7,7) | 0.344 | 0.094 | 0.144 | 0.194 | 0.244 | 0.294 | 0.344 | 0.394 | 0.444 | 0.494 | 0.544 | 0.594 |
| X(8,3) | 0.138 | 0.000 | 0.039 | 0.078 | 0.116 | 0.155 | 0.194 | 0.233 | 0.272 | 0.310 | 0.349 | 0.388 |
| X(8,6) | 0.014 | 0.000 | 0.026 | 0.053 | 0.079 | 0.106 | 0.132 | 0.159 | 0.185 | 0.211 | 0.238 | 0.264 |
| X(9,1) | 0.070 | 0.000 | 0.032 | 0.064 | 0.096 | 0.128 | 0.160 | 0.192 | 0.224 | 0.256 | 0.288 | 0.320 |
| X(9,7) | 0.118 | 0.000 | 0.037 | 0.074 | 0.110 | 0.147 | 0.184 | 0.221 | 0.257 | 0.294 | 0.331 | 0.368 |
| X(10,4) | 0.159 | 0.000 | 0.041 | 0.082 | 0.123 | 0.164 | 0.205 | 0.246 | 0.287 | 0.328 | 0.369 | 0.409 |
| X(10,5) | 0.114 | 0.000 | 0.036 | 0.073 | 0.109 | 0.146 | 0.182 | 0.219 | 0.255 | 0.292 | 0.328 | 0.364 |
| X(11,2) | 0.198 | 0.000 | 0.045 | 0.090 | 0.135 | 0.179 | 0.224 | 0.269 | 0.314 | 0.359 | 0.404 | 0.448 |
| X(11,6) | 0.039 | 0.000 | 0.029 | 0.058 | 0.087 | 0.116 | 0.144 | 0.173 | 0.202 | 0.231 | 0.260 | 0.289 |
| X(12,4) | 0.716 | 0.466 | 0.516 | 0.566 | 0.616 | 0.666 | 0.716 | 0.766 | 0.816 | 0.866 | 0.916 | 0.966 |
| X(12,7) | 0.387 | 0.137 | 0.187 | 0.237 | 0.287 | 0.337 | 0.387 | 0.437 | 0.487 | 0.537 | 0.587 | 0.637 |
| X(13,2) | 0.266 | 0.016 | 0.066 | 0.116 | 0.166 | 0.216 | 0.266 | 0.316 | 0.366 | 0.416 | 0.466 | 0.516 |
| X(13,5) | 0.175 | 0.000 | 0.043 | 0.085 | 0.128 | 0.170 | 0.213 | 0.255 | 0.298 | 0.340 | 0.383 | 0.425 |
| X(14,1) | 0.264 | 0.014 | 0.064 | 0.114 | 0.164 | 0.214 | 0.264 | 0.314 | 0.364 | 0.414 | 0.464 | 0.514 |
| X(14,3) | 0.255 | 0.005 | 0.055 | 0.105 | 0.155 | 0.205 | 0.255 | 0.305 | 0.355 | 0.405 | 0.455 | 0.505 |
| X(15,4) | 0.201 | 0.000 | 0.045 | 0.090 | 0.135 | 0.181 | 0.226 | 0.271 | 0.316 | 0.361 | 0.406 | 0.451 |
| X(15,7) | 0.217 | 0.000 | 0.047 | 0.093 | 0.140 | 0.187 | 0.233 | 0.280 | 0.327 | 0.373 | 0.420 | 0.467 |
| X(16,3) | 0.407 | 0.157 | 0.207 | 0.257 | 0.307 | 0.357 | 0.407 | 0.457 | 0.507 | 0.557 | 0.607 | 0.657 |
| X(16,6) | 0.287 | 0.037 | 0.087 | 0.137 | 0.187 | 0.237 | 0.287 | 0.337 | 0.387 | 0.437 | 0.487 | 0.537 |
| X(17,1) | 0.683 | 0.433 | 0.483 | 0.533 | 0.583 | 0.633 | 0.683 | 0.733 | 0.783 | 0.833 | 0.883 | 0.933 |
| X(17,5) | 0.421 | 0.171 | 0.221 | 0.271 | 0.321 | 0.371 | 0.421 | 0.471 | 0.521 | 0.571 | 0.621 | 0.671 |
| X(18,2) | 0.215 | 0.000 | 0.046 | 0.093 | 0.139 | 0.186 | 0.232 | 0.279 | 0.325 | 0.372 | 0.418 | 0.465 |
| X(18,7) | 0.176 | 0.000 | 0.043 | 0.085 | 0.128 | 0.170 | 0.213 | 0.255 | 0.298 | 0.340 | 0.383 | 0.426 |
| X(19,4) | 0.066 | 0.000 | 0.032 | 0.063 | 0.095 | 0.127 | 0.158 | 0.190 | 0.222 | 0.253 | 0.285 | 0.316 |
| X(19,5) | 0.150 | 0.000 | 0.040 | 0.080 | 0.120 | 0.160 | 0.200 | 0.240 | 0.280 | 0.320 | 0.360 | 0.400 |
| X(19,6) | 0.132 | 0.000 | 0.038 | 0.076 | 0.115 | 0.153 | 0.191 | 0.229 | 0.267 | 0.306 | 0.344 | 0.382 |
| X(20,2) | 0.323 | 0.073 | 0.123 | 0.173 | 0.223 | 0.273 | 0.323 | 0.373 | 0.423 | 0.473 | 0.523 | 0.573 |
| X(20,6) | 0.609 | 0.359 | 0.409 | 0.459 | 0.509 | 0.559 | 0.609 | 0.659 | 0.709 | 0.759 | 0.809 | 0.859 |
| X(21,7) | 0.176 | 0.000 | 0.043 | 0.085 | 0.128 | 0.170 | 0.213 | 0.255 | 0.298 | 0.340 | 0.383 | 0.426 |
| X(21,3) | 0.207 | 0.000 | 0.046 | 0.091 | 0.137 | 0.183 | 0.229 | 0.274 | 0.320 | 0.366 | 0.411 | 0.457 |
| X(22,5) | 0.195 | 0.000 | 0.045 | 0.089 | 0.134 | 0.178 | 0.223 | 0.267 | 0.312 | 0.356 | 0.401 | 0.445 |
| X(22,6) | 0.239 | 0.000 | 0.049 | 0.098 | 0.147 | 0.196 | 0.244 | 0.293 | 0.342 | 0.391 | 0.440 | 0.489 |
| **Mean Abs Dev** | | **0.179** | **0.137** | **0.097** | **0.064** | **0.038** | **0.036** | **0.078** | **0.121** | **0.164** | **0.207** | **0.250** |