An analytical framework for monitoring and optimizing bank branch network efficiency

E.H. Smith
12018570

Dissertation in partial fulfilment of the requirements for the degree Master of Commercii at the Potchefstroom campus of the North-West University

Supervisor: Prof. H.A. Kruger
Co-supervisor: Dr. R. Goede

November 2009

ABSTRACT

Financial institutions make use of a variety of delivery channels for servicing their customers. The primary channel utilised as a means of acquiring new customers and increasing market share is through the retail branch network. The 1990s saw the Internet explosion and with it a threat to branches. The relatively low cost associated with virtual delivery channels made it inevitable for financial institutions to direct their focus towards such new and more cost efficient technologies. By the beginning of the 21st century and with increasing limitations identified in alternative virtual delivery channels, the financial industry returned to a more balanced view which may be seen as the revival of branch networks. The main purpose of this study is to provide a roadmap for financial institutions in managing their branch network. A three step methodology, representative of data mining and management science techniques, will be used to explain relative branch efficiency. The methodology consists of clustering analysis (CA), data envelopment analysis (DEA) and decision tree induction (DTI). CA is applied to data internal to the financial institution for increasing the discriminatory power of DEA. DEA is used to calculate the relevant operating efficiencies of branches deemed homogeneous during CA. Finally, DTI is used to interpret the DEA results and additional data describing the market environment the branch operates in, as well as inquiring into the nature of the relative efficiency of the branch.

Keywords:

Financial industry, data mining, management science techniques, clustering analysis, data envelopment analysis, decision tree induction, homogeneity, positivistic research, quantitative analysis, interpretative research, qualitative analysis.

OPSOMMING

Finansiële instellings maak gebruik van verskillende metodes om dienste aan kliënte te lewer. Die fundamentele metode van dienslewering, werwing van nuwe kliënte en markuitbreiding is die gebruikmaking van die netwerk van takke. Die vinnige ontwikkeling van die Internet in die 1990s het 'n groot bedreiging vir takke ingehou. Alternatiewe elektroniese bankdienste vereis veel laer inset- en operasionele kostes, gevolglik het finansiële instellings hulself op hierdie nuwe tegnologie toegespits. Teen die begin van die 21ste eeu het sekere tekortkominge van die elektroniese bankdienste duidelik geword en het finansiële instellings opnuut die belangrikheid van dienslewerende takke besef. Dit kan as die herlewing van banktak netwerke beskou word. Die hoofdoel van die studie is die daarstelling van 'n wegwyser waarvolgens 'n netwerk van takke bestuur kan word. Die metodiek is verteenwoordigend van data ontginning en bestuurswetenskaplike metodes. Dit stel 'n metodologie voor bestaande uit drie stappe, te wete groeperingsanalise (CA), besluitnemingsbome (DT) en data bekledingsanalise (DEA) ten einde die relatiewe effektiwiteit van takke te verstaan. Tydens die eerste stap word takke deur middel van CA groepeer sodat takke in 'n groep soortgelyk aan mekaar is. Sodoende word die onderskeidingsvermoë van DEA verbeter. In die tweede stap word DEA gebruik om te bepaal watter tak, relatief tot die ander in die groep, die beste van hulpbronne gebruik maak. Laastens word resultate van DEA, tesame met data van die markomgewing waarin die tak bedryf word, ontleed.

Sleutel woorde:

Finansiële instellings, data ontginning, bestuurswetenskaplike metodes, groeperingsanalise, besluitnemingsbome, data bekledingsanalise, positivistiese navorsing, kwantitatiewe ontleding, interpretiewe navorsing, kwalitatiewe ontleding.

Acknowledgements

Writing this dissertation was one of my most fulfilling and proudest life journeys. During this time I have grown immensely, not only moving closer to becoming a master in my field from a theoretical and practical point of view, but also on a personal level. My eyes have been opened to a new world where diligence, discipline and perseverance take a new meaning. This opportunity to grow would not have been possible without the support of many people. My sincerest thanks to all of them in particular the following:

- My Father in heaven, who gave my wisdom, strength and perseverance to successfully complete this study.
- A special word of thanks to Prof. Hennie Kruger, my supervisor, for his encouragement, support and willingness to share knowledge.
- Dr. Roelien Goede, my co-supervisor, for her support and comments.
- Dr. H.R. van der Walt for editing this dissertation.
- The North-West University for making this opportunity possible.
- My parents, Herbie and Alet, and my siblings, Madelein and Herbie, for all their love and support for which I am very grateful.

Table of contents

Chapter 1 - Research synopsis

1.1 Introduction

Financial institutions make use of a variety of delivery channels as a means to liaise and service customers. The primary channel to acquire new customers and increase market share is the branch network. The branch network requires a major capital investment and is committed to expenses with regard to operations and human resources (Cavell *et al.* 2002:63). Regardless of such large investment requirements, branches remain the key to financial institutions' profitability (Cavell, 2002:1). A bank branch may be seen as a location where a financial institution offers a wide variety of face to face services to its customers. Services provided by branches range from cash transactions with a bank teller to financial advice through a specialist. Branches can be situated in a shopping mall or a standalone building.

A branch that operates efficiently offers a multitude of positive possibilities in the form of high customer acquisition and profitability resulting in a favourable return on investment and a gain in competitive advantage over others. In contrast to such branches, those not doing so well can be a huge burden. Inefficient branches will not only be less profitable but capital invested could have been invested elsewhere in more profitable assets and can be seen as a lost opportunity.

According to Cavell *et al.* (2002:69), branches operating efficiently have succeeded in aligning resources to market opportunities. Optimisation of the branch network entails the process of identifying profitable market opportunities, then aligning available resources such as capital, operating expenses and personnel to these identified opportunities. The complicated nature of the subject makes it particularly difficult for financial institutions to identify branches that would operate efficiently and fully understand the reasons for branch efficiency. The difficulty resides in the fact that a financial institution not only needs to understand and manage its resources, but must additionally take into consideration the influence of the area surrounding the branch. Thus, a financial institution must fully comprehend the impact of the market environment on the operations of a branch. Understanding the surrounding area requires the assembly and exploration of factual information concerning the market place and organising it in such a manner as to draw attention to promising opportunities.

This study attempts to establish an analytical model that would assist executives in optimising the network of branches by looking at branch efficiencies from an internal and external point of view. The following section elaborates on the importance of such a study.

1.2 History of bank branches and relevance of the study

The 1990s saw the Internet explosion and with it a threat to branches. Virtual delivery channels, such as Internet banking, compared to physical branches, require a relatively low investment and operational cost. Financial institutions continuously strive towards improving shareholder value and can achieve this by either taking greater risk, or reducing operational cost. The relatively lower cost associated with virtual delivery channels made it inevitable for these institutions to focus on new, more cost efficient, virtual technologies.

In the UK, the move towards virtual delivery channels was of such significance that branches operated by financial institutions were reduced by 25%. In addition to lower operational cost evidence, particularly from the European countries, virtual delivery channels were practically feasible with market penetration up to 29% for telephone banking and 7% for Internet banking. With cost saving opportunities and the possibility of good market penetration, branches came under severe pressure from analysts to abandon them and embrace the virtual age (Cavell, 2002:5).

Despite the fact that analysis illustrated the cost effectiveness of virtual delivery channels, it became evident that such channels failed in sustaining and developing customer relationships. Virtual delivery channels were regarded by customers as merely an addition to the branch and not a complete substitute for the branch. Cameron *et al.* (2006:263) suggest that face to face accessibility will remain important to the majority of customers, while a minority of about 12 per cent of the customer base will handle their finances entirely remotely. Furthermore, compared to branches, low internet costs associated with virtual delivery channels were unable to generate cost to income ratios such as branches (Cavell, 2002:6).

In addition to the short comings of virtual delivery channels, it also became apparent that certain key functions could not be performed by any other channel except branches. Branches are regarded as representative of the brand and serve as the principal channel for customer relationship management and acquisition. Branches also provide critical services for certain types of customers, such as the business sector (Cavell, 2002:6).

In recent years, the financial industry returned to a more balanced view of delivery channels, which may be interpreted as a revival of branch networks. Cavell (2002:6) argues that institutions that had always recognised the importance of branches and continued their branch network development programmes are reaping the benefits of leadership in this field. With a clear understanding of the importance of branches, financial institutions started re-evaluating their branch networks with the emphasis on optimising the current network.

1.3 Objectives of this study

The primary objective of this study is to establish an analytical model that would assist executives of a financial institution in optimising the network of branches. A combination of data mining techniques, i.e. clustering analysis (CA) and decision tree induction (DTI), together with a management science technique known as data envelopment analysis (DEA) will be utilised to exploit financial and geo-demographic data. The outcome of this study will be used to better understand factors that drive branch efficiency and in doing so, eventually raise their performance to a higher level.

A second application of this study is aimed at branch network maintenance by means of benchmarking branches within the network. According to Cavell (2002:160), benchmarking, also known as comparative analysis, has been adopted by many organisations as a means of identifying and implementing best practice by measuring performance internally or against competitors. This will enable managers to identify branches not operating efficiently and pinpoint features that, if improved, will elevate the branch to a higher level of efficiency and ultimately higher profitability.

1.4 Limitations of this study

In this study, the operating efficiency of branches of a financial institution is investigated with the focus on providing management with data driven knowledge that would assist them in decision making regarding the branch network. Therefore, it is reasonable to limit it to the use of generic data mining and management science techniques and exclude attempts to create or improve analytical methods.

Acknowledging the fact of stiff competition within the financial industry, this study was restricted to focusing on the branch network of a single financial institution. The methodology utilised can be applied to a variety of industries however the data and results obtained cannot automatically be applied to other financial institutions.

Acceptance testing of the analytical model presented in this study is constrained by the fact that measuring the influence of a practical implementation requires a prolonged study. Therefore, findings were presented to management for a subjective interpretation based on domain knowledge.

1.5 Research methodology

The main theory underlying the proposed framework is to study the characteristics of branch efficiency within the branch network and then derive rules as to the magnitude and combination of features essential to have a branch operating efficiently. This study can be divided into a literature study section and an empirical study section.

1.5.1 Literature study

A literature study was conducted to determine the feasibility of creating an analytical framework that can assist financial institutions in optimising the branch network. The literature study performed can be divided into two parts. The first part (Chapter 2) focuses on the selection of a research methodology and the applicability of selected analytical methods to this particular study. It may be seen as the roadmap for this study. The second part (Chapter 3) of the literature research focuses on the theory relating to the analytical methods used.

Vassiloglou and Giokas (1990:591) conducted a study at the Commercial Bank of Greece in assessing the relative efficiency of bank branches using DEA. Characteristics analysed were labour, supplies example stationery, monetary value of branch installation, number of computer terminals and transactions processed. Results presented to top management, were found to correspond to examinations previously done by management.

Sherman and Ladino (1995:60) applied DEA to a major bank with 33 branches and used it to identify ways to improve productivity within the bank. The results provided the basis for reviewing and evaluating branch operations. Using the obtained results the bank was able to substantially improve branch productivity and profits. Implementing changes in branch operations as a result of DEA lead to annual savings of over $6 million.

Fatti and Clarke (1999:57) conducted a study at a major South African bank to determine the manpower requirements needed, utilising DEA to identify efficient branches. In this particular study, number of employees as input variables and average volumes of work as output produced by the branch were analysed. The application of that study resulted in a 7.5 % saving of total manpower requirements.

It became evident that revival of branches is receiving a great deal of focus, furthermore DEA, the method used for selecting efficient branches in many instances, proofed to be very effective.

The proposed analytical method deviates from the above-mentioned studies in a number of aspects. This framework attempts to combine data internal and external to the financial institution. Internal data reflect factors the financial institution can manipulate, for example amount paid for rent and total human resource expense. Data external to the financial institution present factors it has little or no control over and do not specifically relate to the financial institution. The study is based on the market environment the branch operates in and provides additional information to this end. Relevant factors consist of geo-demographic characteristics, for example literacy of population, average population and current community size. Internal data will be used to derive branch efficiency, while external data will serve to identify the market environment needed for a branch to be efficient.

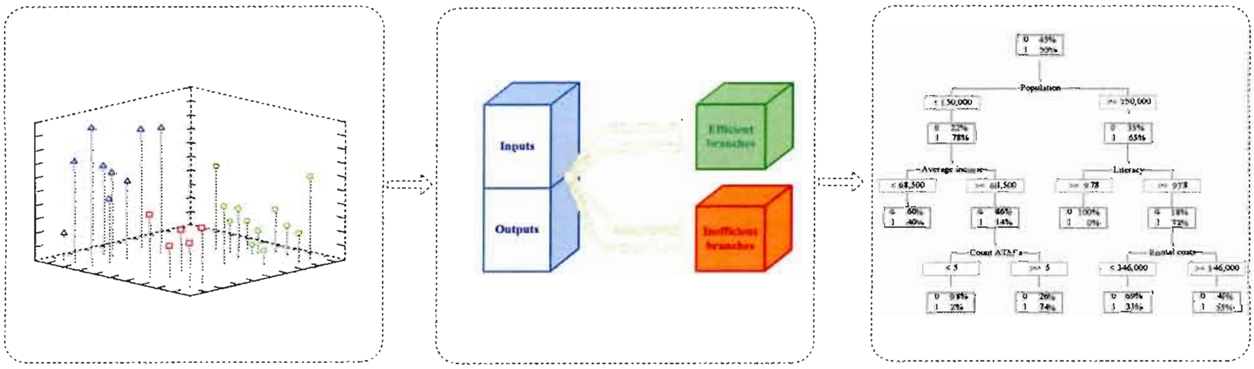
Secondly, this framework will, similar to the previously mentioned studies, exploit the effectiveness of DEA. However, data used as input for DEA will be homogeneous with regard to specific characteristics. This will be accomplished by using *k*-means clustering. In Section 2.4.2 (Chapter 2), the importance of working with homogeneous branches is highlighted. The following section briefly describes the empirical study conducted.

1.5.2 Empirical study

Identifying efficient branches encompass comparing branches within the branch network of the financial institution the study is done on, thus enabling the model to differentiate between efficient and inefficient branches. Vast differences such as demographic and household income levels to name but two, necessitate as a first step the identification of homogeneous branches. It is important to compare homogeneous branches as results will be inaccurate if a branch from a highly urbanised area, such as Sandton, is compared to a branch from a rural area such as Pongola. Homogeneous branches, for comparison purposes, will be grouped (clustered) together by using the *k*-means clustering algorithm. A motivation for the selection of analysis methods will be given in Chapter 2.

Branches clustered together as homogeneous will be evaluated to differentiate between efficient and inefficient branches. A special type of linear programming application known as DEA will be applied to the homogeneous branches for labelling a branch as efficient or inefficient. DEA was used to determine which of the decision making units (DMUs), in this case branches of a financial institution, make efficient use of resources available to them. Motivation for the use of DEA is given in Chapter 2.

After evaluating the efficiency of branches, a decision tree (DT) is built with the efficiency, determined by the DEA of every branch, as the target variable. Finally, the rules generated by DT are used to elucidate on key factors impacting on branch operating efficiency. The analytical model consists of three steps. Figure 1.1 outlines the overall process of this study.



Step1: Clustering analysis

Step2: Data envelopment analysis

Step3: Decision tree induction

Figure 1.1 Graphical illustration of proposed analytical model.

Processes depicted in Figure 1.1 are highly dependent on one another in order to produce high quality results. For example, if in industrial mining the process of filtering waste and precious minerals is inadequate, all subsequent processes will be affected. The following section briefly covers the Chapter layout of the study.

1.6 Dissertation layout

Chapter 2 discusses various research methodologies. Special attention will be given to the importance of using a specific methodology within its philosophical framework, as well as advantages and disadvantages associated with the use of research methodologies. Literature study relates to methods proposed in this analytical model

In Chapter 3, the theory of data mining and management science techniques used in this study is discussed. The aim of this chapter is to give an overview of the analytical methods used, as well as alternative methods available. Motivation for the selection of the methods used will be elaborated on.

Subsequent chapters report on the empirical study as follows:

Chapter 4 describes the application of CA to obtain homogeneous branches.

Chapter 5 describes the use of DEA to identify branches that operate efficiently. In addition, benchmarking of branches to pinpoint inefficiencies associated with inefficient branches is also covered.

Chapter 6 contains the application of a DT to DEA results. The chapter focuses on explaining the DEA results obtained in Chapter 5, by looking at data internal to the financial institution and data describing the geo-demographic environment.

The thesis concludes with a summary of the study given in Chapter 7 and results obtained by the analytical model. In addition, the chapter also highlights limitations of the study for feature purposes.

1.7 Chapter conclusion

Chapter 1 presented the history of branches and described how branches of financial institutions could come under severe pressure of alternative delivery methods, such as the Internet. Through a good understanding of the history of branches, the usefulness and application of this study become evident. The remainder of this study describes the methodology used to aid managers in maintaining the branch network.

Chapter 2 - Research methodology

2.1 Introduction

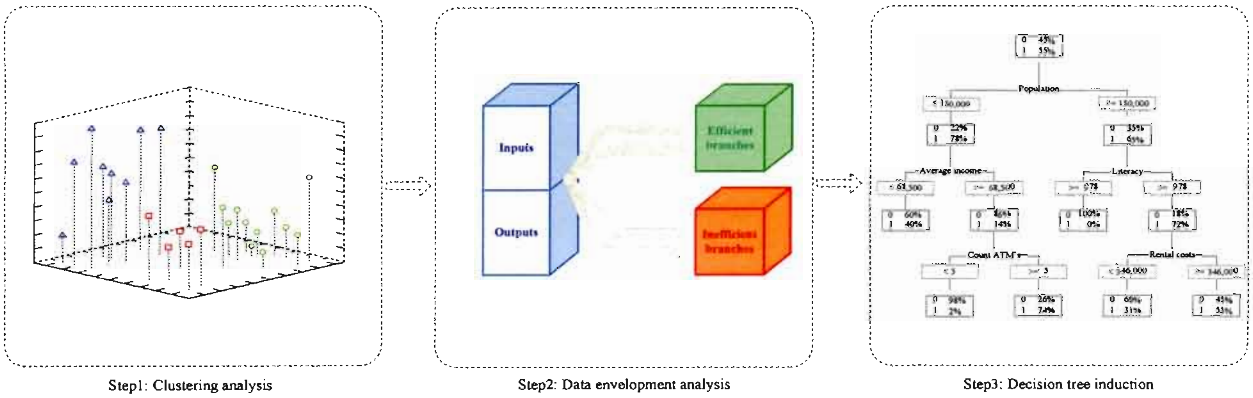
The importance for a financial institution to properly manage its branch network was discussed in Chapter 1. The urgency of managing a network of branches received particular attention since the threat that the Internet will replace physical branches. Fortunately this threat never materialised but it brought to the attention of financial institutions the significant part a branch plays in the financial institution. With a clear understanding of this necessity, objectives were defined in Chapter 1 to create an analytical model that would aid financial institutions in managing the network of their branches.

Chapter 2 starts with a synopsis of the proposed analytical model followed in Section 2.3 covering various research methodologies available to the researcher. Research methodology selection may be a challenging task for the researcher. With methodology superiority being a favourite topic often found in research literature, selecting an appropriate research methodology may be difficult (Karami *et al.,* 2006:43). For this reason a noticeable portion of this chapter will be devoted to the topic of research methodology.

The remainder of Chapter 2 is separated into quantitative (Section 2.4) and qualitative (Section 2.5) research methods. In these sections, a literature study covering research philosophies underlying these methodologies, as well as analytical methods used in this study, relating to either of these groups are covered. In addition, Chapter 2 also reports on previous studies where these methods were applied. Chapter 2 concludes with Section 2.6 on the ability of mixing different research methods and a brief overview of research classifications in Section 2.7. Chapter 3, the second part of the literature study, will elaborate on the technical theory relating to the methods applied in this study.

2.2 Synopsis of proposed analytical model

This research proposes a three step analytical model to assist management of a financial institution with the management of the branch network. As previously mentioned, this would be an ex ante tool, that would be to aid the financial institution when faced with selecting between a number of alternatives. Figure 2.1 shows the proposed model with the first step being clustering analysis (CA) followed by data envelopment analysis (DEA) and the final step of decision tree induction (DTI).



Step1: Clustering analysis

Step2: Data envelopment analysis

Step3: Decision tree induction

Figure 2.1 Graphical illustration of analytical model as ex ante tool to aid financial institutions in managing the branch network.

CA is applied in the first step of the analytical model to enforce a high level of homogeneity during DEA. A high level of similarity is important to ensure that DEA results are not biased towards larger branches in any way. The second step in the analytical model, DEA, identifies which branches of the financial institution operate efficiently. The final step of the model uses the results created by DEA, additional variables describing the geo-demographical environment the branch operates in and a decision tree (DT) to extract rules explaining branch efficiencies.

Samoilenko and Osei-Bryson (2008) proposed the use of a similar model in a study conducted on countries that were transitioning from centralised to market economies. Kumar and Ravi (2007) conducted a literature study of statistical and intelligent techniques that have been used in the financial sector between 1968 and 2005. They list many studies where several analytical techniques have been combined to improve bankruptcy prediction accuracy.

2.3 Research methodologies

Research is a universal activity by which a specific phenomenon is studied objectively in order to create a valid theory explaining the topic (Fox & Bayat, 2007:4). Various research methodologies are available for conducting research and selecting an appropriate method may be challenging.

Research methodologies and research approaches on a high level get classified as being either quantitative or qualitative of nature. Quantitative research methods most often refer to positivistic methods whereas qualitative research methods usually refer to the interpretive, social constructionism and subjective methods (Maree, 2007:50). Fox and Bayat (2007:65) assert that the quantitative and qualitative research paradigms differ extensively, principally as a result of the philosophical assumptions on which they are based. It would be unwise to study a topic in a particular manner without being aware of the philosophical backgrounds underlying the research paradigms.

The importance of understanding fundamental differences between research philosophies is summarised by Easterby-Smith *et al.* (2002:27) as follows:

- A thorough understanding of research philosophies will simplify research design. Research design involves identifying methods for collecting and investigating collected evidence;
- Knowledge about research designs will make it clear to the researcher which research designs are applicable to which problems, therefore avoiding possible research design errors;
- A broad understanding of research philosophies enables the researcher to apply research methods previously unknown to him/her.

The following sections discuss the quantitative and qualitative classification of research methods and also describe the philosophy underlying the research approach. Research methods relating to either a quantitative or qualitative research approach used in this study will be reviewed.

2.4 Quantitative research approach

Positivistic research methods which are mostly quantitative in nature dominated the research science as the preferable research approach for the greater part of the 20th century (Maree, 2007:50). A key feature when conducting research from a positivistic viewpoint is the fact that the research is conducted in an objective manner. The researcher will collect data in a value free manner, thus not showing any form of biasness during the data collection phase, and objectively interpret the collected data allowing the data to speak for itself (Saunders *et al.*, 2000:85).

Research conducted within the positivistic paradigm focuses on research using a highly structured approach in order to aid in the reproduction of quantifiable observations that lead to the analysis. According to Karami *et al.* (2006:48), popular techniques utilised during quantitative research include analysis techniques such as factor analysis, correlation analysis, cluster analysis, regression analysis and non-parametric analysis methods.

Quantitative research is firmly rooted in the positivistic research philosophy. The next section gives a brief outline of the positivistic research philosophy.

2.4.1 Philosophical background to positivistic research methods

Positivism originated from the natural science and according to Brewerton and Millward (2001:11), positivism implies an approach to research where problems are objectively investigated in order to derive rules and laws describing the phenomena.

The positivistic approach to research requires an environment where the researcher has complete control over variables in order to measure their effects on other variables (Maree, 2007:55). Important to note is that even though the researcher has complete control over the experiment, the researcher objectively participates and records findings quantitatively to conclude laws that govern phenomena researched.

Easterby-Smith *et al.* (2002:28) encapsulate the core characteristics of positivism as the following:

- independence: the observer must be independent from what is being observed;
- value-freedom: the choice of what to study, and how to study it, may be determined by objective criteria rather than by human beliefs and interests;
- hypothesis and deduction: science proceeds through a process of hypothesis sizing fundamental laws and then deducing what kinds of observations will demonstrate the truth or falsity of these hypotheses;
- operationalisation: concepts need to be operationalised in a way which enables facts to be measured quantitatively;
- reductionism: problems as a whole are better understood if they are reduced into the simplest possible elements;

- generalisation: in order to be able to generalise about regularities in human and social behaviour, it is necessary to select samples of sufficient size from which inferences may be drawn about the wider population;
- cross-sectional analysis: such regularities can most easily be identified by making comparisons of variations across samples.

Positivists assert that scientific methods produce precise, verifiable, systematic and theoretical answers to the research questions (Maree, 2007:55). The fact that these methods provide quantifiable results can be seen as a benefit of using research methods relating to the positivistic approach. For example, customer satisfaction at branch A can be expressed as 80% satisfaction compared to customer satisfaction at branch B which is 95%. Reiteration of the research by another researcher, barring that variables stay constant, will produce the same results. Another useful property of the positivistic methods is that research can periodically be performed and the amount of change can be measured.

On the negative side, the positivistic approach is heavily criticised for reducing problems to purely numbers. Prasad (2005:6) emphasises that positivism is somewhat inadequate for the understanding and investigation of subtle differences as those associated with complex real world processes. During positivistic research, problems often get simplified, as real world problems are much too complicated to be fully expressed numerically. The argument is that a lot of detail is lost during the problem reduction phase. Another drawback of this approach is the question of how to accurately express the feelings and emotions of people as numbers while maintaining objectivity.

2.4.2 Positivistic methods used in this research

Affordability of modern technologies has made it possible for companies to have databases containing terabytes of data (Berry & Linoff, 2004:476; Cabena *et al.*, 1998:9; Hadjinian *et al.*, 1998:9; Hormozi & Giles, 2004:62). In fact, huge data volumes may be seen as the norm. Data captured in databases contain real life information describing daily business events and are generally stored at an atomic level. Since data is not randomly generated, it represents actual customer needs and preferences and is rich with valuable, hidden information that can be used for making well informed business decisions.

Modern day profusion of data resulted in data abundance as opposed to a paucity of data in the past. Difficulty associated with data abundance is not a hardware related problem in the sense of processing power shortage. Enormous volumes of data available for analysis make it difficult, nearly impossible, for humans to understand and comprehend. To resolve the problem of extracting information from large quantities of raw data, managers depend on models that discern raw data and highlight recursive patterns. Hormozi and Giles (2004:63) claim that prior to analysis techniques such as data mining, managers were not as capable of making informed decisions, since searching through enormous amounts of data was too expensive and time-consuming and in some instances even impossible.

This study uses a combination of data mining and management science techniques for the creation of an analytical model that will assist management with managing the institution's branch network. Managing the branch network in the context of this research refers to tasks such as adding new branches to the existing branch network and identifying branches that do not operate efficiently. Figure 2.1 in Section 2.2 is a graphical representation of the analytical model proposed for this research. The first and last steps in the model, CA and DTI, are well known data mining techniques, while the second step uses a management science technique known as DEA.

Even though data mining and management science techniques share a single goal, i.e. to assist management in decision making by providing information, they have different backgrounds and are exploited for solving different kinds of problems. The boundaries of these techniques are increasingly overlapping, and Berry and Linoff (2000:18) predict that in future, techniques used to extract knowledge from data will be integrated.

The following sections briefly explore data mining and management science techniques as positivistic methods used in this study. A literature study relating to the use of these decision aiding techniques will be evaluated.

2.4.2.1 Background to data mining

For some time, businesses have been searching for methods that would allow them to gain insight from massive amounts of generated data. This initial search to gain insight led to the development of online analytical processing (OLAP) tools. These tools were good for all-purpose reporting but fell short of producing insights required by modern businesses. Research spotlight therefore shifted to data mining with the focus of utilising large business data bases (D' Souza *et al.*, 2007:281).

Data mining surfaced during the late 1980's and may be seen as a multidisciplinary field utilising techniques from a diversity of disciplines (Han & Kamber, 2006:5). These disciplines include research areas such as database technology, machine learning, statistical pattern recognition, neural networks and artificial intelligence, to name but a few. Data mining as a discipline can be seen as a result of the natural evolution of information technology. A key characteristic of data mining is the fact that these algorithms have been developed to cope with large quantities of data (Ramakrishnan & Gehrke, 2003:890).

The general idea behind data mining, from a customer relationship point of view, is to scrutinise masses of raw data describing customer behaviour for hidden information. Hormozi and Giles (2004:62) reference several definitions for data mining as listed in Table 2.1. Data mining is regarded as a hypothesis-free approach, in other words searching for previously unknown patterns, in contrast to most popular statistical methods requiring the development of a hypothesis in advance (Cabena *et al.*, 1998:17). This is also a second key characteristic of data mining.

Authors	**Popular definitions of data mining**
Berry and Linoff (2000)	Data mining is "the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules."
Hui and Jha (1999)	"Data mining is the process of discovering interesting knowledge from large amounts of data that may be used to help companies make better decisions and remain competitive in the marketplace."
Chung and Gray (1999)	"The objective of data mining is to identify valid, novel, potentially useful, and understandable correlations and patterns in existing data."
Fabris (1998)	Data mining is described as the automated analysis of large amounts of data to find patterns and trends that may have otherwise gone undiscovered.
Cabena *et al.* (1998)	Data mining is defined as "the process of extracting previously unknown, valid, and actionable information from large databases and then using the information to make crucial business decisions."
Fayyad *et al.* (1996a)	Data mining is a step in the knowledge discovery in databases (KDD) process and refers to algorithms that are applied to extract patterns from the data. The extracted information can then be used to form a prediction or classification model, identify trends and associations, refine an existing model, or provide a summary of the database being mined.

Table 2.1 Data mining definitions found in literature.

Data mining assists businesses to better understand their business hence enabling the business to better serve the customers (Chopoorian *et al.*, 2001:45). Many industries employ data mining as it has proven itself a valuable tool. Industries utilising these methods range from the financial sector, retail and insurance industries, telecommunications to the health sector.

Financial institutions in particular generate large quantities of data and by using data mining for analysis of patterns and trends, financial institutions have been able to accurately predict, for example, how customers will respond to interest rate increases, as well as identifying high risk customers likely to default a loan (Hormozi & Giles, 2004:63).

The following is an example of the resourcefulness of data mining. The South African Police Service's Crime Information Analysis Centre reported that 225 bank robberies took place during the first 6 months of 2001. Loss as a result of robberies for that period totalled a staggering 27 million Rand with ABSA, South Africa's largest bank group, being targeted more than any other bank in South Africa. Bank robberies are not only dangerous, but also harmful to a bank's image, which should represent a safe environment for keeping one's valuables. Through the use of data mining and spatial mapping, ABSA managed to create a profile of branches likely to be robbed and then placed precautionary security measures at such high risk branches. SAS Institute (2008) states that with a 38 % reduction in cash lost and a reduction of 41 % in armed robberies, this system proved to be very effective.

Berry and Linoff (2000:8) separate data mining tasks in two groups, directed and undirected mining techniques. The difference between techniques from these categories is that directed mining techniques try to predict the value of a desired variable, whereas undirected methods try to find hidden patterns without the use of a target variable.

Berry and Linoff (2000:8) describe directed mining models as methods attempting to predict the value of a particular target attribute, such as income or response. These models are trained, using a training dataset where the target field is known and is therefore called supervised learning. Such models can predict for example, the likelihood that a customer will respond to a direct marketing campaign, the probability that a customer might swap to a competitor (also called churn) and fraud in the credit and debit card industry.

Undirected mining models on the other hand attempt to find patterns among groups of records without the use of a target attribute. This is known as unsupervised learning. Undirected mining models are often used during the data exploration phase to identify new patterns or segmentations. As there are no pre-classified data, it is up to a subject master to determine whether the patterns have any significance. Descriptive models consist of two fundamental model types, clustering and association models. Clustering (also called segmentation) groups similar people, objects or events together in a cluster. Also by grouping similar objects together in clusters, these clusters now represent all the objects in the respective clusters.

2.4.2.2 Data mining techniques used in this study

This study uses CA and DTI from the data mining area. The following sections cover a literature review of the data mining methods mentioned.

2.4.2.2.1 Clustering analysis to identify homogeneous branches

This study compares branches of a financial institution to determine which of the branches make efficient use of given resources. DEA is employed to determine which of the branches operate efficiently. The section covering management science techniques used in this study will elaborate on efficiency analysis. This section reports on a literature study done on similarity grouping, also known as cluster analysis (CA).

Similarity grouping is the event of grouping similar objects together and is particularly important in this research. According to Samoilenko and Osei-Bryson (2008:1577), results gained from DEA can be improved when objects compared have a high level of similarity. Fatti and Clarke (1998) also grouped similar branch types together in their study where their aim was to predict human resource requirements in a bank environment. Since results gained from DEA can be improved in the presence of homogeneity, similarity grouping will be employed as the first phase of this analytical model. Figure 2.2 shows cluster analysis as a pre-process to DEA. In the diagram it can be seen that the triangles are separated from the squares and circle figures. In the same sense will certain branches be separated from others, based on certain criteria, for example, the number of accounts and number of customers, *et cetera*.

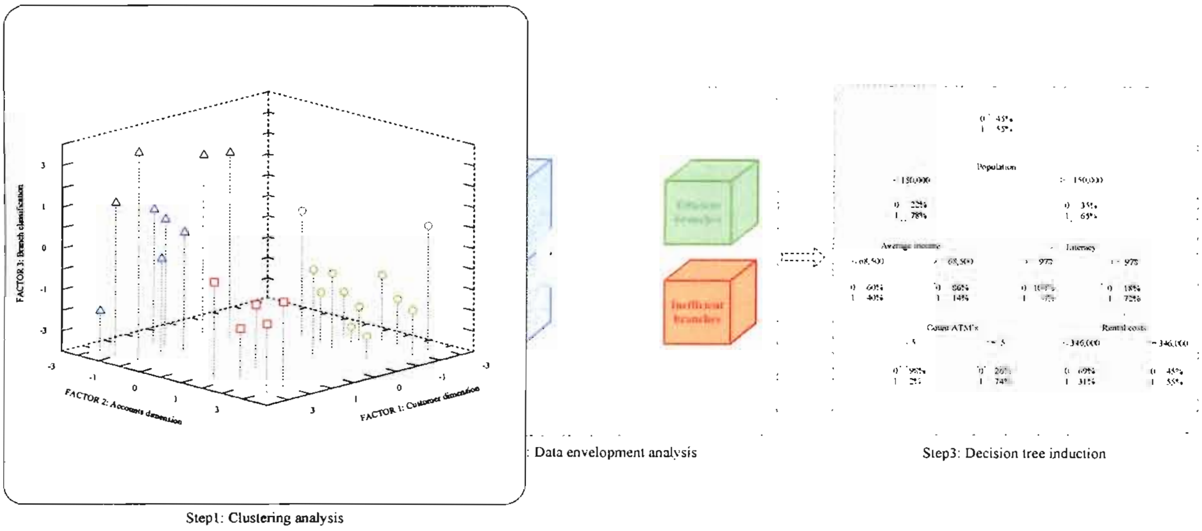

FACTOR 3: Branch classification
FACTOR 2: Accounts dimension
FACTOR 1: Customer dimension
Step1: Clustering analysis
Inefficient branches
: Data envelopment analysis
Population
< 150,000
>= 150,000
Average income
Literacy
Count ATM's
Rental costs
Step3: Decision tree induction

Figure 2.2 CA as first step in the process.

CA is a popular undirected data mining technique that is used to identify homogeneous objects (Kononenko & Kukar, 2007:13). Large volumes of literature available on the topic of similarity grouping serve as confirmation of the usefulness and extensive popularity of this particular research field. In addition to similarity grouping, CA is also a popular method for subset selection. Working with a larger dataset occasionally requires the selection of a subset of data representing the original population. In these cases, clustering is a known method to aid in the subset selection procedure (Daszykowski *et al.*, 2002:91).

Customer relationship management is an area where CA is commonly used (Ngai *et al.*, 2009:2592). It is also frequently applied to problems in research areas such as econometrics (Kim & Ahn, 2008; Ye, 2007; Murphy *et al.*, 2007; Díez *et al.*, 2006), financial sector (Cameron *et al.*, 2006; Sinka & Corne, 2004; Mills, 2000) process analysis (Ahvenlampi & Kortela, 2005), psychological research (Stefurak & Calhoun, 2007), sociological research (Wang & Zaïane, 2002) and medical research (Lin & Chien, 2009; Sewitch *et al.*, 2004).

A practical example of CA done at Experian Business Strategies, a financial marketing company, highlights the effectiveness of cluster analysis (Cameron *et al.*, 2006). During 2004 the UK's financial industry spent £723 million on direct mail. The amount spent on direct mail in the financial industry, is noticeably higher than the amount spent by other market sectors. If the selection criteria used to identify customers to target via mail could be enhanced, it would result in significant savings for companies in the financial sector. A new customer segmentation model was developed, taking into account individual and household attributes of a customer and therefore looking at a customer holistically. A *k*-means clustering technique was used to cluster customers that were described by more than 350 measures. The new solution had an $18\% - 21\%$ increase in accuracy when compared to the previous model. CA has been applied to many similarity grouping problems with great success. The next section states arguments for the use of CA.

2.4.2.2.2 Applicability of clustering to this research

Branches of a financial institution differ extensively in services they offer and customers they serve. Branches located in rural areas would offer vastly different services to a particularly different customer base than a branch located in a highly populated, commercialised area. Therefore, in order for DEA to produce meaningful results, a high level of homogeneity within branches compared is essential. For that reason the application of CA to branches being studied in this research is of the utmost importance.

Similar branches of a financial institution can be identified by asking an expert within the financial institution that has a thorough understanding of the current branch network. Creating groups, clusters, containing similar branches with ten or even twenty branches to choose from would be a daunting, but possible task. However, the expert would only be able to make logical groupings. Clusters identified by expert opinion would largely be subjective to his/her personal interpretation of the individual branches under investigation. Substantiating similarities found within clusters in the absence of quantitative measures describing similarity, as in the case with expert opinion, will be difficult.

Jain *et al.* (1999:268) argue that humans can, without a doubt, perform competitively against automatic clustering algorithms in a two dimensional space, however, they assert that with the increase in the dimensionality of the problem, so does the difficulty levels of intuitively interpreting the data increase. With well over four hundred branches and fifteen attributes describing each individual branch it becomes an impossible task for even the most seasoned expert to resolve without the use of a clustering algorithm. Identified inadequacies associated with expert intervention for identifying homogeneous branches necessitate exploration of various alternatives. In a different research study Sinka and Corne (2004:132) studied the classification of bank documents. In that particular research they argued for the use of a standard data set to be used for all experiments conducted during their research. They argued that the use of different data sets would require human intervention, for interpretational purposes, and stated that it would be too time consuming and complex for humans to perform.

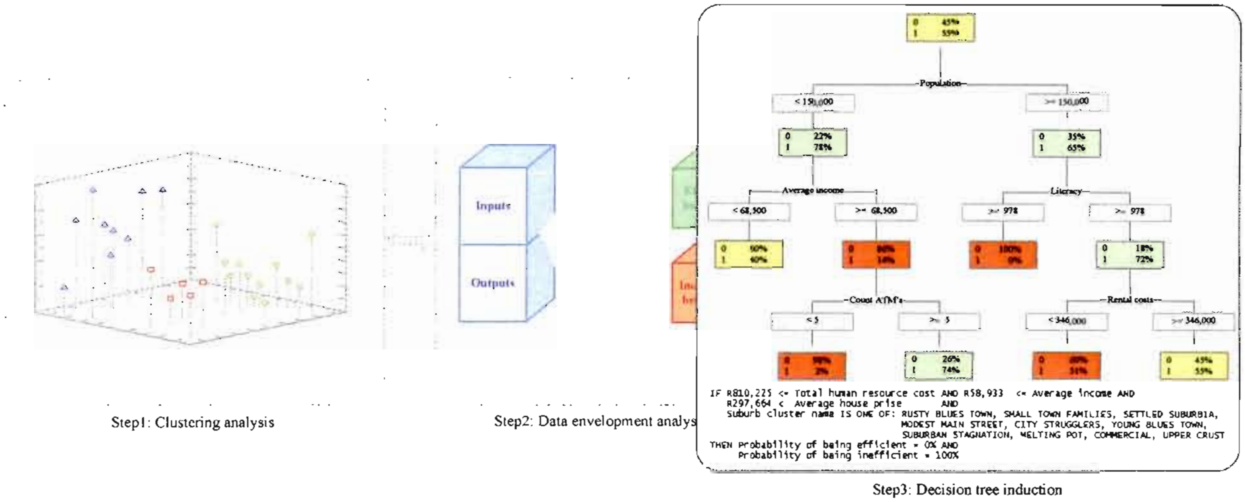
Jain *et al.* (1999) and Sinka and Corne (2004) expressed the fact that human interpretation is inadequate for similarity grouping of objects with high dimensionality. From this it can be concluded that the use of an automated method to perform similarity grouping in this study will be essential. In the previous section covering literature study, it became clear that CA has successfully been applied in many areas for the function of similarity grouping. On that account, CA will be used for grouping similar branches together as a first phase of this research in order to improve DEA results. Chapter 3 will cover CA from a technical perspective.

The following section covers the literature study conducted on the second data mining technique used as part of this study.

2.4.2.2.3 Derive business rules explaining reasons for branch efficiency

Ultimately the goal of this study is to declare reasons to management as to why some branches of the financial institution operate efficiently. DEA will be used in the second step of the analytical model to determine which branches, comparing similar branches, operate efficiently. DTI will be put to use in the third and final step of the analytical model, utilising output generated from DEA, as input for the DT to extract reasons why branches of the financial institution operate efficiently. See Figure 2.3 for a graphical representation of the third step. DEA is categorised as a management science technique and is discussed in sections 2.4.3.

DTI is a powerful and popular directed data mining technique used to build predictive models (Berry & Linoff, 2004:165). The wide popularity of this method is due largely to the fact that DTs represent rules that can easily be expressed as business rules management can interpret, for example, "If branch has *x* customers, *y* accounts and *z* transactions, then it will be efficient".



```
IF R810,225 <= Total human resource cost AND R58,933  <= Average income AND
   R297,664 <  Average house prise       AND
   Suburb cluster name IS ONE OF: RUSTY BLUES TOWN, SMALL TOWN FAMILIES, SETTLED SUBURBIA,
                                  MODEST MAIN STREET, CITY STRUGGLERS, YOUNG BLUES TOWN,
                                  SUBURBAN STAGNATION, MELTING POT, COMMERCIAL, UPPER CRUST
THEN Probability of being efficient = 0% AND
     Probability of being inefficient = 100%
```

Figure 2.3 Graphical representation of the third step in the analytical model, DTI to obtain business rules explaining reasons for branch efficiency.

Proof of the effectiveness of DTI can be seen in the following example. Evans and Fisher (1994) applied DTI at the largest printing company in the US with great success. The DT was, unlike other applications of DT not used to predict outcomes but to provide implementable guidelines that will minimise the occurrence of cylinder banding. An incident of cylinder banding is recognised as a streak of ink running across the printed image ruining the print job. The implementation of a DT significantly reduced the down time account of cylinder banding, thus proving to be a very effective solution (Evans & Fisher, 1994:66).

Decision trees have been used copiously to aid management in decision making (D' Souza *et al.*, 2007:282). Other examples where DTs have been used include econometrics (Kim *et al.*, 2001; D'Souza *et al.*, 2007), the financial sector (Lu & Chen, 2009; Sun & Li, 2008; Florez-Lopez, 2007), process analysis (Sohn & Moon, 2004; Evans & Fisher, 1994; Watkins *et al.*, 2006) and medical research (Porcel *et al.*, 2008; Kunene & Weistroffer, 2008). Kumar and Ravi (2007) conducted a literature study of statistical and intelligent techniques used to solve the bankruptcy prediction problem. Their study emphasises the ability of DTs to create human comprehensible "if – then" rules as a major advantage associated with DTs. Figure 2.3 illustrates a graphical representation of a DT and also displays an example of an "if – then" rule created by a DT.

2.4.2.2.4 Applicability of decision tree induction to this research

The second step in the analytical model, DEA, (step 2 in Figure 2.1), determines and labels branches as operating efficiently or inefficiently, based on preset criteria. For a useful interpretation of DEA results, it is necessary to combine the information with whatever other information is available relevant to branches of the financial institution.

Difficulties anticipated with the third step of the analytical model are that a multitude of factors exist describing the market environment and geo-demographical factors, while selecting attributes that best describe market environment may be difficult. DTI was chosen in this study to combine DEA results with additional information and create rules in easily understandable terms explaining branch efficiency. DTI was selected as tool as it is often applied to derive business rules explaining associates between data and expressing them in such terms. In addition, DTs are a favourite data exploration tool assisting in the identification of important variables and do not require prior domain knowledge or parameter settings (Lu & Chen, 2009:3538). Technical details concerning DTI will be covered in Chapter 3.

2.4.3 Management science techniques used in this study

This study uses a management science technique known as DEA to distinguish between branches of the financial institution that operate efficiently and those that do not. This section starts by giving a short overview of management science techniques and two convincing examples of the effectiveness of these analytical methods. Following this section, a brief literature overview of DEA is presented.

Quantitative and scientific techniques were first developed to assist the military with decision making during World War II (Render & Stair, 2000:3). Utilisation of these newly developed techniques proved to be very useful. Companies which recognised the value of these techniques started applying similar techniques, with great success, to assist with managerial decision making and planning in the corporate environment. These techniques became known as management science techniques and are also referred to as quantitative analysis techniques.

Whitten *et al.* (2001:46) define management science techniques as an information system that provides users with decision-oriented information. These systems can be seen as the scientific approach – which excludes emotions, whim and guesswork – to managerial decision making

(Render & Stair, 2000:2). Management science techniques require the processing and manipulation of raw data into meaningful information.

Nearly any imaginable problem has been successfully addressed with the aid of management science techniques and numerous application examples of these techniques can be found in the literature. As an indication of the resourcefulness of these techniques, two examples will be discussed briefly.

In the early 1990s, North Carolina was spending $150 million on transporting students to schools. The state's transporting system consisted of more than 13,000 buses, 700,000 students and 100 school districts. A decision support system was developed to reduce the costs associated with the transporting system (Sexton *et al.*, 1994). Between 1990 and 1993, through the use of a management science technique, the state saved $25.2 million in capital costs and $27.9 million in operating costs.

Delta Airlines took advantage of management science techniques called "Coldstart" to save an immense $220,000 per day (Subramanian *et al.*, 1994). Coldstart is an exceptionally arduous decision support system and is run daily at Delta Airlines. The objective of this management science system is to minimise operating costs and lost passenger revenue. The system consists of 40,000 constraints and 60,000 variables. A constraint is a restriction on available resources for Delta Airlines. These resources include for example aircraft availability, balancing arrivals and departures at airports and aircraft maintenance needs. Delta Airlines estimated a saving of over $300 million over the course of three years since the management science technique became operational. The following section covers a brief literature study concerning DEA.

2.4.3.1 Differentiation between efficient and inefficient branches

Identification and differentiation of efficient branches can be seen as the core of the analytical model. Branches of the financial institution identified as homogeneous with regard to certain dimensions, will be compared using DEA, to reveal efficient and inefficient branches.

DEA combines inputs and outputs with regard to a certain decision making unit (DMU) into a single efficiency score relative to other DMUs in the study. A DMU in this study will relate to branches of a financial institution. Results from the DEA analysis will then be used as a target variable during the third step, the rule induction phase, of the model. Figure 2.3 illustrates DEA as the second step in the analytical process. This section covers a brief literature study on the subject of efficiency rating, and three real world implementations of DEA will be elaborated on.

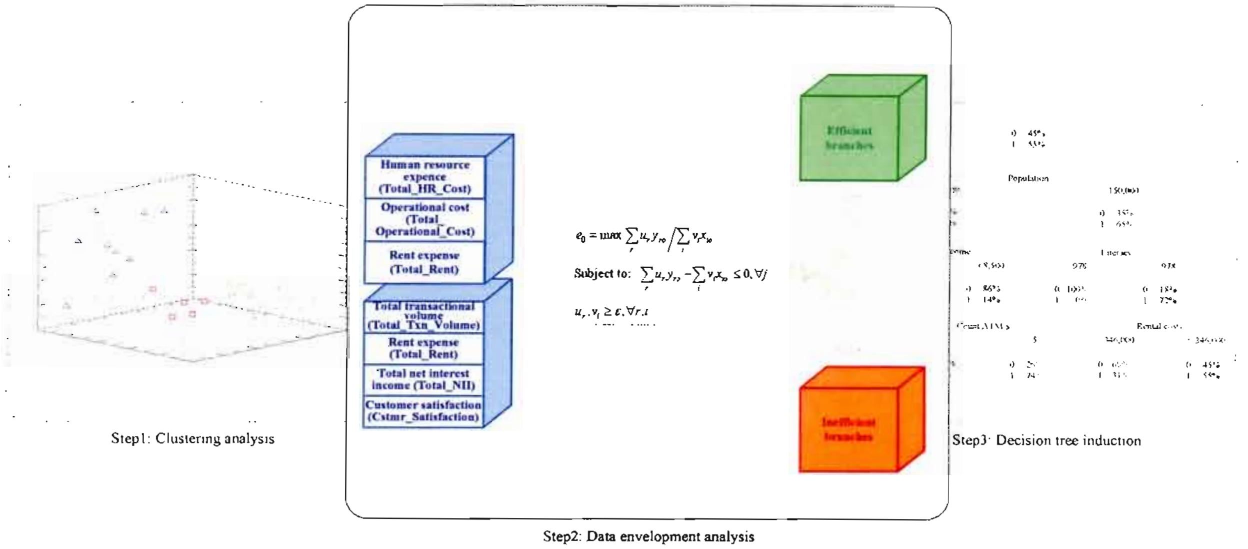

Step1: Clustering analysis
Human resource expence (Total_HR_Cost)
Operational cost (Total_Operational_Cost)
Rent expense (Total_Rent)
Total transactional volume (Total_Txn_Volume)
Rent expense (Total_Rent)
Total net interest income (Total_NII)
Customer satisfaction (Cstmr_Satisfaction)
$e_0 = \max \sum_r u_r y_{ro} / \sum_i v_i x_{io}$
Subject to: $\sum_r u_r y_{rj} - \sum_i v_i x_{ij} \le 0, \forall j$
$u_r, v_i \ge \varepsilon, \forall r, i$
Efficient branches
Inefficient branches
Step2: Data envelopment analysis
Population
Step3: Decision tree induction

Figure 2.3 DEA methods to distinguish efficient branches of the financial institution from inefficient branches.

DEA has been used in many applications and in a diversity of situations but none more than in the financial sector, especially the banking sector. Table 2.2 gives a comprehensive overview of previous research where DEA was actuated to assist management in decision making. A number of completed research studies, listed in Table 2.2, are referenced by Mostafa (2008:310).

DEA is rapidly gaining status as the leading tool for determining efficient operating units, and its popularity can be seen in the growth of real world applications of this technique (Vassiloglou & Giokas, 1990:591).

DEA was first applied to the banking sector by Sherman and Gold (1985). The motive for the research was purely to experiment whether DEA could successfully be applied to the banking sector. Deficiencies associated with traditional financial ratios as a measure of efficiency and the increasing pressure to improve banking performance can be seen as the main driving force for the research.

Study	Country	N	Inputs	Outputs
Sakar (2006)	Turkey	11	Branch numbers, employees per branch, assets, loans, deposits	ROA, ROE, interest income/assets, interest income/operating income, non-interest income/assets
Wu *et al.* (2006)	Canada	142	Employees, expenses	Deposits, revenues, loans
Howland and Rowse (2006)	Canada	162	Non-sales FTE, sales FTE, size, city employment rate	Loans, deposits, average number of products/customer, customer loyalty
Ho and Zhu (2004)	Taiwan	41	Capital stocks, assets, number of branches, employees	Sales, deposits
Mukherjee et al. (2002)	India	68	Net worth, borrowings, operating expenses, employees, number of branches	Deposits, net profit, advances, non-interest income, interest income
Seiford and Zhu (1999)	USA	55	Employees, assets, capital stock	Revenue, profits
Golany and Storbeck (1999)	USA	182	Employees, space, marketing	Loans, deposits, accounts per customer, satisfaction
Drake and Howcroft (1999)	UK	250	Number of loan accounts, number of mortgage accounts, number of cheque accounts	Personal loans, new cheque accounts, mortgage loans, insurance commission, change in 'marketed alliances'
Zenios *et al.* (1999)	Cyprus	144	Employees, terminals, space, current accounts, savings accounts, credit applications	Number of transactions
Ayadi *et al.* (1998)	Nigeria	10	Interest on deposits, expenses on personnel, total deposits	Total loans, interest income, non-interest income
Al-Shammari and Salimi (1998)	Jordan	16	Selected financial ratios	Selected financial ratios
Chen and Yeh (1998)	Taiwan	34	Employees, assets, number of branches, operating costs, interest expenses	Loans, investments interest income, non-interest income
Athanassopoulos (1997)	Greece	68	Employees, ATMs, terminals, interest costs, non-interest costs, location	Non-interest income
Resti (1997)	Italy	270	Employees, capital Loans, deposits, non-interest	Income
Bhattacharya *et al.* (1997)	India	74	Interest expense, operating expense	Advances, deposits, investments
Schaffnit *et al.* (1997)	Canada	291	Employees	Transactions, maintenance
Athanassopoulos and Curram (1996)	UK	250	ATMs, employees, counter transactions, potential market, loans	Sales, liability sales, investments and insurance policies sold
Sherman and Ladino (1995)	USA	33	Employees, expenses, rent	Number of transactions
Favero and Papi (1995)	Italy	174	Employees, capital, loan able funds, deposits, loans, investment in securities, non-interest	Income
Al-Faraj *et al.* (1993)	Saudi Arabia	15	Employees, location, expenses, acquired equipment	Net profit, balance of current accounts, savings account, loans, number of accounts
Fukuyama (1993)	Japan	143	Employees, capital, funds from customers	Loan revenue, other revenues
Giokas (1991)	Greece	17	Employees, expenses, rent	Number of transactions
Oral and Yolalan (1990)	Turkey	20	Employees, terminals, number of accounts, credit applications	Number of transactions

Study	Country	N	Inputs	Outputs
Vassiloglou and Giokas (1990)	Greece	20	Employees, suppliers, space, Computer terminals	Number of transactions
Parkan (1987)	Canada	35	Employees, expenses, space, rent, terminals	Number of transactions, customer response, correction of errors
Sherman and Gold (1985)	USA	14	Employees, expenses, space	Number of transactions

Table 2.2 Extended use of DEA for modelling financial institutions (Mostafa, 2008:310).

Sherman and Gold (1985:297) state that traditionally, banks endeavoured marketing campaigns to promote new products, and processes aimed at improving cash management to enhance bank performance. Very little time was spent on auditing process efficiencies as a means of improving bank performance. In the case of Citicorp, a 1% decrease in operating expenses in 1982 resulted in a 2% increase in net income and earnings per share.

In view of limitations identified with traditional measures by Sherman and Gold (1985:299), they sought different alternatives to measure financial branch efficiency. DEA was proposed as a promising alternative. DEA was applied to 17 savings bank branches located in the same metropolitan area to determine the feasibility of this technique. Table 2.3 lists the inputs and outputs identified by Sherman and Gold as measures for bank branch efficiency.

Inputs (resources used)	Outputs (results produced by consuming inputs)
Labour – full time personnel per branch	Number of transactions – grouped into four groups of similar resource intensity
Rent – as representative to space	
Total cost of office supplies used	

Table 2.3 Inputs and outputs used by Sherman and Gold (1985) in their study of branch efficiency.

Results produced by the DEA study conducted by Sherman and Gold (1985) led to the following conclusions. A branch identified by DEA as inefficient was already marked by management for termination. Furthermore, an additional four branches identified as inefficient by DEA were previously under investigation, and it was concluded that branch inefficiency was a result of managerial weaknesses. Therefore, results obtained by DEA were consistent with management's perceptions. Besides previously mentioned correlations, DEA recognised branches believed to be efficient by management, as inefficient. Thus, for this bank, DEA provided additional insight that could lead to additional efficiency and profitability.

Initial scepticism of management about the usefulness of DEA results was astonishing, as they requested new branches acquired by the bank to be measured against existing branches to determine the efficiency of the new branches. Enclosed, DEA is very capable of determining the efficiency at which branches operate.

Human resource related costs constitute a major part of banks' expenses and serve as a measure used to determine efficiency with DEA in the second example. A study conducted by Fatti and Clarke (1998) at Standard Bank of South Africa attempted to forecast human resource requirements over a long term.

Various factors were taken into consideration when estimating human resource requirements. The most significant requirements were the volume of work done per branch. In addition, some non-clerical roles needed regardless of the workload also had to be considered. In order to determine the human resource requirements, the expected workload over the longer term would have to be predicted, and resources required to process the predicted workload calculated. The branches of the bank were grouped into 5 homogenous groups depending on the type of work they performed. DEA was then separately applied to every group to identify the efficient branches. Table 2.4 lists the variables that were used in the research as part of DEA.

Inputs (employment grades)	**Outputs (volumes of work)**
Branch manager	No. of cheque accounts with credit balances
Branch administrator	No. of cheque accounts with debit balances
Clerical grades	No. of savings accounts
Checking grades	No. of cheque account paper-based debit entries
Supervisory grades	No. of cheque account paper-based credit entries
	No. of savings account paper vouchers
	No. of teller transactions
	No. of foreign exchange transactions

Table 2.4 Input and output variables used during research by Fatti and Clarke (1998) at Standard bank of South Africa.

Once efficient branches per group were identified, regression models for every category gave promising results. As a final step in predicting human resource requirements over the longer term, future workloads had to be forecast. A sensible amount of historical data was required for forecasting techniques, such as seasonal smoothing and ARIMA modelling. The project team developed a custom forecasting model. This model was fitted to all branches in the five categories to forecast future workload. The newly developed model was tested by applying the fitted regression models to the actual workloads over the previous six months. These results were compared to the results obtained from the bank's heuristic method for selecting efficient branches. The new method resulted in a 7.5 % lower human resource requirement than the previous model. Since the only difference between the new and the previous model is the way in which efficient branches are selected, it is clear from this second example that DEA is an excellent technique for selecting efficient units.

In the third example, efficiency at which branches of the Commercial Bank of Greece operate, was evaluated. This strive towards improved decision making was mainly driven by the emerging financial integration of the European Economic Community (EEC) countries. Vassiloglou and Giokas (1990) calculated branch efficiency by using DEA. Table 2.5 lists the attributes used in the research done at the Greek bank.

Inputs (resources used)	**Outputs (grouped by level of difficulty)**
Labour: impact on cost and production	Group A
Supplies used by branches: level of production	Group B
Branch installation: measured in terms of floor space	Group C
Number of computer terminals	Group D

Table 2.5 Variables used in the application of DEA to the Commercial Bank of Greece.

Results obtained from this study were compared to data already available to the bank. It was agreed that results obtained through this model were generally in line with previous results. Results in conflict with information previously obtained, could be accounted for as a lack of data availability.

This section covered three examples where DEA was intelligently and successfully applied in the financial industry. Despite the fact that DEA is a relatively new management science technique, many more compelling examples exist in literature.

2.4.3.2 Applicability of data envelopment analysis in this research

Bank performance is usually measured using accounting ratios such as return on assets and return on investment. While these measures present adequate information, they suffer from several drawbacks. Limitations identified in accounting ratios are (Sherman & Gold, 1985:299):

- Accounting measures aggregate many aspects of branch office management, for example marketing and operations. A branch might perform exceptionally well in a specific aspect of branch operations and fairly poor in another and owing to aggregation might still be presented as performing well;
- Branch offices might process many basic, non-fund generating transactions. The efficiency of these branches might appear to be lower than other branch offices but these branches provided basic daily services required by customers;
- Particular transactions are more resource intensive, for example issuing a bond, and may negatively influence a branch office's performance, since a branch with a high mixture of resource intensive transactions might appear to be relatively inefficient;
- Many traditional accounting measures do not evaluate how efficiently resources or inputs were utilised in order to achieve services or transactions.

O'Donnell and Van Der Westhuizen (2002:86) assert that standard financial measures like return on assets, return on equity, etc. have another short coming. When determining branches operating efficiently, they have to be compared to a standard or benchmark while obtaining a suitable benchmark may be difficult. Premachandra *et al.* (2007:414) record additional advantages of DEA when comparing against alternative statistical and economical methods.

DEA has been identified as a superior technique when faced with the problem of differentiating between efficient and inefficient decision making units. This technique has frequently been applied to the banking industry, as the three examples in the previous section proved. On account of examples listed, DEA was selected as analytical method to differentiate between efficient and inefficient branches.

The capability of DEA to pinpoint inefficiencies, as well as the magnitude of inefficiency associated with a decision making unit can be seen as an additional benefit of using this management science technique. DEA will also be used as a benchmarking tool to improve inefficient branches as benchmarking of branches is the second objective of this study. A technical discussion of DEA models follows in Chapter 3.

2.5 Qualitative research approach

Interpretive research, which is primarily qualitative by nature, focuses on human attitudes, behaviour and experiences and tries to ascertain an in-depth opinion of a particular group of people (Dawson, 2007:15). During qualitative research, fewer participants take part in the research but contact with the participants is a lot more detailed, and contact sessions with participants last longer. Qualitative research is usually described as methods trying to collect descriptive rich data and developing an understanding of what is being studied through focusing on how individuals and groups view and interpret the world. Even though this research fully relied on quantitative methods, interviews were scheduled with relevant parties to discuss results obtained.

Brewerton and Millward (2001:12) list the following characteristics that distinguish qualitative, interpretive and quantitative, positivistic research methods:

- focus on interpretation rather than quantification;
- an emphasis on subjectivity rather than objectivity;
- flexibility in the process of conducting research;
- an orientation towards process rather than outcome;
- a concern with context - regarding behaviour and situation as inextricably linked in forming experience;
- an explicit recognition of the impact of research process on the research situation.

According to Maree (2007:58) interpretivism is the oldest philosophical strand qualitative research links to. The next section gives an overhead perspective of how the interpretive philosophical background influenced qualitative research methods.

2.5.1 Philosophical background to interpretive research methods

Interpretive traditions materialised from a scholarly position that trades human interpretation as the starting point for developing knowledge about the social world (Prasad, 2005:6). During the last part of the 19th century, philosophers formulated a new paradigm called social constructionism. Easterby-Smith *et al.* (2002:28) also refer to social constructionism research philosophy as interpretive methods.

Easterby-Smith *et al.* (2002:28) state that social constructionism and therefore the interpretive research philosophy was developed due to the deficiencies identified when applying a positivistic research methodology. Lee (1991:342) claims that the interpretive approach to research has been gaining attention in recent years as an alternative to the traditional positivist approach. The interpretive perspective is based on the following key assumptions (Maree, 2007:60):

- Human life may only be understood from within. It cannot be observed from some external reality. Interpretivism therefore focuses on people's subjective experiences, on how people "construct" the social world by sharing meanings, and how they interact with or relate to each other.
- Social life is a distinctively human product. Interpretivism assumes that reality is not objectively determined, but is socially constructed. The underlying assumption is that by placing people in their social contexts, there is a greater opportunity to understand the perceptions they have of their own activities.
- The human mind is the purposive source or origin of meaning. Through uncovering how meanings are constructed, we can gain insights into the meanings imparted and thereby improve our comprehension of the whole.
- Human behaviour is affected by knowledge of the social world. Interpretivism proposes that there are multiple and not single realities of phenomena and that these realities can differ across time and place.
- The social world does not "exist" independently of human knowledge. As researchers, our own knowledge and understanding of phenomena constantly influence us in terms of the types of questions we ask and the way we conduct our research. Our knowledge and understanding are always limited to the things to which we have been exposed, our own unique experiences and the meanings we have imparted.

The main advantage of the interpretive research philosophy according to Goede and Kruger (2004:8) is the fact that the researcher has a holistic view of the problem. This observation rises from the fact that the researcher has the opportunity to interpret answers from questionnaires against a background gained from interacting with respondents. Thus, by understanding the background of the problem domain, the researcher might arrive at a more enlightened conclusion than what would have been the case had the researcher only relied on numbers. On the other hand, in the absence of numerical facts, managers used to working with numbers, might find it hard to compare cyclically research results. This characteristic of interpretive research philosophy can then also be seen as the main weakness when applying interpretive methods.

2.6 Making use of multiple research methods

In some instances it might be beneficial to combine research methods from different backgrounds, and it is quite usual for a single research study to combine quantitative and qualitative methods (Saunders *et al.*, 2000:98). The combination of various research methods is known as triangulation (Dawson, 2007:22). Saunders *et al.* (2000:99) list the following benefits associated with using multiple research methods:

- Different methods may be used for different purposes in a study. You may wish to employ case study methods, for example interviews, in order to get a feel for the key issues before embarking on a survey. This would give you confidence that you were addressing the most important issues;
- using a multi-method approach enables triangulation to take place. Triangulation refers to the use of different data collection methods within one study in order to ensure that the data are telling you what you think they are telling you.

Research methodologies do not exist in isolation and research methodologies frequently utilise techniques that can be traced back to different philosophies. Therefore, in some instances, it may be difficult to clearly link a research method to a research philosophy. For example, in order to better understand financial institution branch efficiency, the researcher might interview branch managers to better grasp factors associated with branch expenses. Factors that surfaced during interviews as important might then be used in empirical methods instead of using empirical methods for determining variable importance. In this example, it would have been beneficial to use a multi-method approach.

2.7 Research classifications

Research can be classified depending on the expected outcome or intended use of the research. According to Easterby-Smith *et al.* (2002:8), three main categories of research exist; applied, pure or action research.

When the intention of the research is to solve a specific problem it is classified as applied research. Usually in the case of applied research, clients are confronted with a problem and are willing to pay for a solution to their problem (Easterby-Smith *et al.,* 2002:9). Although theory is used during problem solving, it is the application of theory that is important in this case. Applied research usually forms the foundation of Masters Degrees where students describe what they have done, for the purpose of the clients, and explain why they used selected methods, for academic purposes. The research done in this study falls within the applied research category by using a conglomeration of three known methods to assist a financial institution in selecting and maintaining the branch network. This serves as an ex ante tool in aid of choosing among different alternatives.

Key characteristics of pure research are that results are openly distributed in books, are mainly for academic audiences and are mainly intended for theoretical development (Easterby-Smith *et al.* (2002:9). Theoretical development can take on any of the following forms. Theoretical development can be the discovery of new ideas generated from empirical research or invention, where a new technique is developed to deal with a specific problem and in this case have considerable commercial potential. The last form of theoretical development is reflection, where an existing solution is re-examined from a different perspective and used mainly for doctoral theses.

Lastly, research can be classified as action research when the researcher actively participates in the research environment and no longer tries to maintain a distance from the research project. Thus, the researcher and the researched can both be seen as participants in the research process. In the case of action research, researchers try to understand a problem by making changes to a process and understanding the effect of the changes.

2.8 Chapter conclusion

Chapter 2 started with a synopsis of the proposed analytical model from Chapter 1. Figure 2.1 visually displayed the model and this graphical representation of the model will continuously be used and referenced in this research. Research methodologies available, as well as reasons why understanding research methodology can be beneficial to the researcher were expressed.

Section 2.4 gives background knowledge pertaining to the quantitative research methodology. This section also covered a brief literature study on the methods used in this research relating to the quantitative background. The literature study covered CA and DTI. From the literature study, it became clear that both methods from a data mining discipline were best suited for the task at hand. DEA from a management science background were selected as the method for identifying operating units, in this study, branches of a financial institution that operate efficiently given certain inputs to produce outputs. The literature study covering DEA has made it clear that this method was far superior in solving these types of problems.

Although no formal use will be made of qualitative research methods, Section 2.5 covered this research methodology. It was felt that even though these methods were not employed regular contact sessions were held with representatives from the financial institution to discuss results. For that reason Section 2.5 and Section 2.6 were dedicated to qualitative research methods and the possibility of combining these methods.

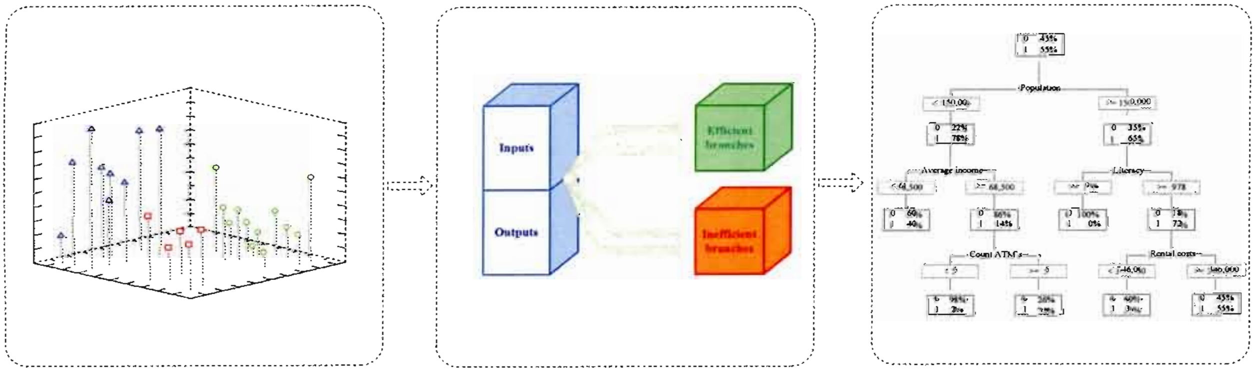
The chapter was concluded with research classifications. The next chapter will cover some of the technical details relating to the methods discussed in this chapter.

Chapter 3 - Theoretical introduction to methods used

3.1 Introduction

The aim of this chapter is to give a theoretical overview of the analytical methods used in this model. Literature study conducted in the previous chapter assessed the applicability of clustering analysis (CA), data envelopment analysis (DEA) and decision tree induction (DTI) to the research objectives at hand. From the literature study it became palpable that the methods and more specifically a combination thereof, offer several benefits when applied to similar problems. Based on these findings, chapter three will cover each of the analytical methods in further detail.

Figure 3.1 is a graphical representation of the analytical model consisting of three distinct steps. The first step is to identify similar branches as it is essential for the remainder of the framework to compare apples with apples. This will be accomplished through the use of CA, an unsupervised data mining technique. The second of three steps and backbone of the methodology (DEA) is logical and discriminates between branches operating efficiently and those that do not.



Step1. Clustering analysis

Step2: Data envelopment analysis

Step3: Decision tree induction

Figure 3.1 Graphical overview of analytical method.

The third and final step of the methodology uses results obtained from DEA, branding a branch as efficient or inefficient, as input to a decision tree (DT). The efficiency label assigned to a branch by means of DEA will be used as target variable during DTI. In addition to the efficiency target variable, additional attributes, mostly attributes describing the external market environment, will be included in the DTI phase.

Internal data refer to data describing characteristics specifically relating to the financial institution the study is based on, for example, management selection and staff compliment. External data describe virtues of the market environment the financial institution operates in and has very little, if any leverage on. Vassiloglou and Giokas (1990:592) refer to these elements as controllable and non-controllable variables and stress the importance of identifying to which type a variable belongs. It would not make sense to identify areas the financial institution has to improve on if it has no means of changing that specific aspect. Data describing the market environment include factors such as average population and average house price, etc. Cluster analysis and DEA utilised internal data, with the exception of customer count that was classified as external data. Inclusion of customer count during the clustering process was necessary to ensure a more accurate identification of similar branches. A comprehensive description of all the data used will be supplied in the subsequent chapters.

A DT is constructed to predict the probability of a potential new branch operating efficiently given a certain market environment. The DT can then be used to derive easy understandable rules to aid during the expansion phase of the branch network. The DT may be seen as the link between internal and external data, in that the internal data determine the efficiency of a branch and the external data describe the market environment needed for such a branch to flourish in, or the inverse, to falter in.

Chapter 3 is divided into three sections. The first section will investigate technical details related to the use of clustering analysis (CA). DEA and DTI will be dealt with in the remaining two sections. In these sections, alternative methods will be addressed and advantages, as well as disadvantages related to the use of these methods, will be highlighted.

3.2 Clustering analysis

Clustering is a method used to group homogeneous objects together and is defined by Jain *et al.* (1999:268) as the separation of a heterogeneous data set into groups of data, each group called a cluster, so that objects within a cluster are homogeneous and different from objects in other clusters. Inter-group homogeneity is achieved by maximising inter-group similarity and minimising intra-group similarity. Holistically, a cluster may be seen as a representation of the objects contained within the cluster. In the case of this research, branches of the same financial institution will be clustered so that branches within a cluster are homogeneous with respect to some similarity criteria. The similarity criteria will be elucidated in Chapter 4. Clustering is a prominent technique that is frequently used for the scientific exploration of data and one of the best techniques to group similar objects together (Liao & Wen, 2007:4).

Several techniques exist to aid in the task of similarity grouping of objects. As mentioned in Chapter 2, an alternative would be to seek the advice of a subject specialist but for reasons already provided, it became evident that a more scientific measure is required. The following two sections on clustering briefly discuss self-organising maps (SOMs), an artificial neural network approach to clustering and alternative clustering algorithms. Particular attention will be given to the *k*-means clustering algorithm, as this clustering method will be applied in this analytical model.

The discussion on clustering methods is followed by a basic comparison between SOMs and the *k*-means clustering methods. The section on clustering closes with an overview on the topic of evaluation of cluster results.

3.2.1 Self-organising maps as clustering method

Self-organising maps is a variant of artificial neural networks and is frequently used during data mining as a tool to group similar objects. Also known as Kohonon self-organising maps, named after the inventor, this method relies on neural networks for CA (Han & Kamber, 2006:434). SOMs are single layer feed-forward neural networks. Feed-forward networks associate input with outputs directly.

Self-organising maps are analogous to neural networks and consist of an input and output layer. Every neuron in the SOM has an independent weight associated with each incoming connection and each unit in the input layer is connected to one source. Differences highlighted by Berry and Linoff (2004:249) between SOMs and neural networks are differences in topology. The former consist of many units – arranged like a grid – instead of just a handful as in the case of a neural network.

Self-organising maps, like neural networks, need to be trained before use. Learning the SOM enables it to detect patterns. Training data, where the value of the target variable is known beforehand, is used to teach the network. Berry and Linoff (2004:249) describe the training process as follows: A case from the training set flows through the network from the input units to the output units. Units in the output layer compete against each other and the unit with the highest value wins. This will be the unit that best correlates with the input pattern presented. The winning unit's weight gets adjusted in order to improve the ability of that unit to recognise similar patterns in the future. In addition to adjusting the weight of the winning unit, units in the immediate neighbourhood are also adjusted. The immediate neighbourhood is determined using the neighbourliness parameter. This parameter controls the size of the immediate neighbourhood, as well as the sternness of the adjustment. Severity of adjustments, to winning and neighbouring units, decreases as training continues (Berry & Linoff, 2004:250).

In general, self-organising maps identify fewer clusters than it has units in the output layer. Validation data, i.e. data that was not used during the training phase, is usually used to validate the SOM. Berry and Linoff (2004:251) argue that validating SOMs serves two purposes. Firstly, it identifies units in the output layer not utilised and which can be removed from the self-organising map. Removing redundant units enhances a SOM's performance, since unnecessary units will not be present to compete for a case presented to the network, and therefore reduce time used by the SOM to identify clusters. Secondly, it identifies whether over fitting occurred during the training phase. Over fitting will be observed if the ability of the SOM to discover patterns in the validation set is inferior to the detection ability illustrated during the training phase. Only once the SOM is in place and tested against the validation set, can it be applied to unseen data.

Self-organising maps, as clustering tool, is highly regarded on account of its ability to handle imprecise and fuzzy information (Liao & Wen, 2007:2). On the negative side, SOMs produce clusters with similar objects but do not give a clear indication as to which attributes within the resulting clusters played a significant part in the creation of the clusters. SOM is defined by its units and weights assigned to units, making networks created enormously complex and interpretation of results awkward. Another weak point identified with SOM is that there is no assurance that optimal weights are assigned to units during the training phase that can lead to patterns in the data being missed (Berry & Linoff, 2004:250).

3.2.2 Clustering algorithms

A large number of clustering algorithms are available and reported on in the literature (Gelbard *et al.*, 2007:156). New algorithms are introduced and older algorithms updated, with every algorithm performing better in its own right when presented with appropriate data. Appropriate data refer to the fact that the natural structure of the data corresponds to the characteristics of the algorithm.

Clustering algorithms generally are grouped into hierarchical or partitioning algorithms. Many different clustering algorithms utilising different methods make categorising algorithms tricky, since some algorithms have properties that are common to various categories (Berkhin, 2002:5). As accuracy of the final result greatly depends on the clustering method selected (Mingoti & Lima, 2006:1742), the rest of this section will focus on different clustering methods.

3.2.2.1 Hierarchical clustering algorithms

Hierarchical clustering algorithms build cluster hierarchies and create clusters in one of two ways, namely an agglomerative or a divisive manner. The agglomerative methods follow a bottom-up approach and the divisive hierarchical algorithms, the invert of the agglomerative methods, a top-down approach when clustering objects. Figure 3:2 illustrates the difference between the agglomerative and divisive hierarchical clustering. Gelbard *et al.* (2007:157) claim that hierarchical clustering algorithms differ mainly in the way distance is measured between clusters.

Agglomerative hierarchical algorithms are usually used as a data explanatory tool and in some cases as a way of determining the natural number of clusters in a data set (Mingoti & Lima, 2006:1745). Clustering starts by assigning each object to a cluster so that each object has its own cluster. Clusters are then merged with the nearest cluster based on a similarity measure. This merging of neighbouring clusters is repeated until the desired number of clusters is obtained. Whilst iterating, only two clusters can be joined, and once two clusters have been merged they cannot be separated again. Agglomerative nesting (AGNES) is a well known agglomerative clustering algorithm.

Divisive clustering starts with all objects contained within one big cluster. This cluster is then repetitively divided into appropriate smaller clusters until the desired number of clusters is obtained. Divisive analysis (DIANA) is an example of a divisive hierarchical clustering method.

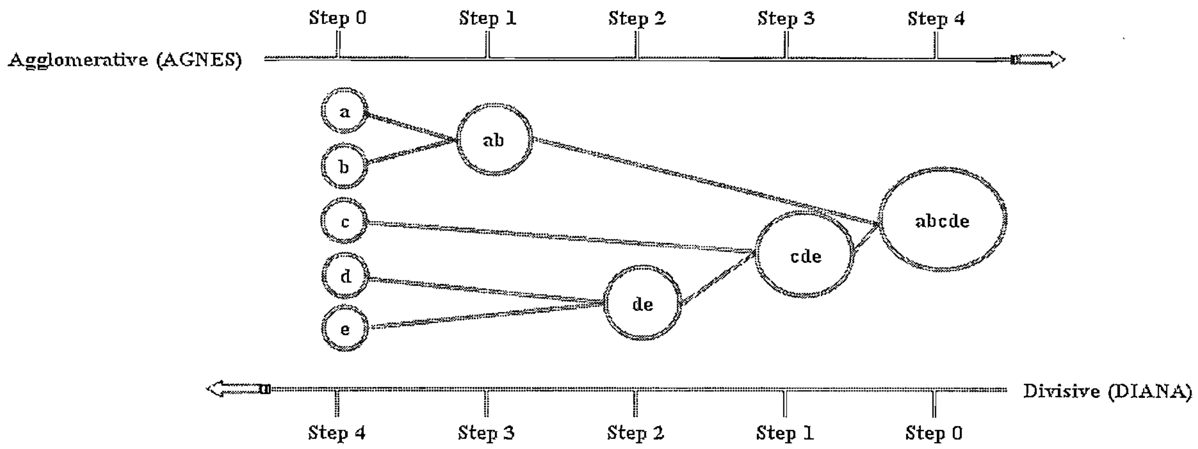


Figure 3.2 Agglomerative and divisive hierarchical clustering of data objects as illustrated by Han and Kamber (2006:409).

Hierarchical clustering allows exploration of data on different levels of detail, and this property makes hierarchical algorithms an attractive tool for clustering similar objects. Han and Kamber (2006:408) draw attention to the fact that pure hierarchical clustering algorithms never return to a cluster once it has been created. The repercussion of never revising created clusters increases the possibility of inferior cluster quality, since hierarchical algorithms never revaluate and change clusters once created, even if a split or merger of clusters was a bad choice.

3.2.2.2 Partitioning clustering methods

Hierarchical clustering algorithms build clusters progressively. Partitioning algorithms on the other hand start by dividing data into several, preliminary, subsets where every subset embodies a cluster. Finding an optimal solution entails testing all possible combinations of objects to cluster assignments. Testing every possible combination would be impractical therefore these methods instead employ heuristic methods, such as *k*-means or k-medoids algorithms, to assist with clustering. Partitioning clustering methods work well for finding spherical-shaped clusters in small to medium sized databases, and when the natural structure of the data corresponds to these algorithms high quality clustering results can be expected (Berkhin, 2002:5).

3.2.2.3 *K*-means clustering

K-means clustering is one of the most popular clustering algorithms used to group similar objects together (Liao & Wen, 2007:4; Mingoti & Lima, 2006:1746). The *k*-means algorithm groups a multi-dimensional data set into a predetermined number of groups called clusters. Objects within a cluster represent similar characteristics and dissimilar characteristics to objects in different clusters.

Macqueen (1967:283) describes the process of *k*-means clustering as follows: The number of clusters - the *k*-value - is specified prior to clustering, hence the name *k*-means clustering. Different *k*-values will return different results and a *k*-value corresponding to the natural structure of the data will generate good clusters. On the other hand, if the value does not match the natural data structure results will be poor.

Once the user specified the appropriate *k*-value, the algorithm starts by selecting *k* arbitrary data points as the initial cluster centres, also called centroids. As the initial cluster centres are selected randomly, they may cause a variation in the groupings obtained on successive runs. Randomly selecting cluster centres, to start with, do not appear to be a serious concern. Minor shifts, on consecutive runs, are due to the unavoidable difficulty that some points are located between clusters (Macqueen, 1967:290). All data points are then assigned to the nearest arbitrary selected cluster centre and can only belong to one cluster. Subsequently, a new mean is calculated for every cluster and this becomes the new cluster centre or centroid. All the data points are then reassigned to the nearest centroid. The mean for every cluster is then re-calculated. At any point the *k*-means are actually the arithmetic mean of the groups of clusters it represents. This is an iterative process and continues until data points stop moving between clusters. Figure 3.3 is a graphical illustration of the *k*-means clustering process. With reference to the successive iterations, observe the cluster centre moving as objects get assigned to new clusters.

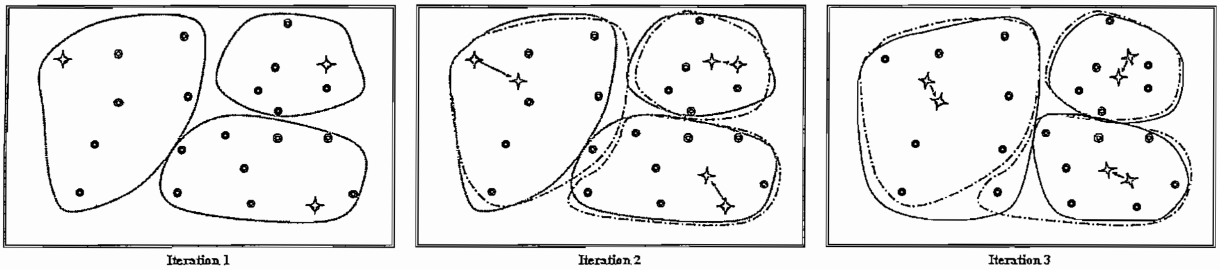


Figure 3.3 Clustering of a set of objects with the *k*-means method as illustrated by Han and Kamber (2006:403).

Analysing every possible subset of clusters is computationally infeasible and various greedy heuristics are used for iteration optimisation. The squared error clustering algorithm is generally applied, and this algorithm minimises the squared error. Dunham (2003:140) describes the squared error of a cluster as the sum of the squared Euclidean distances between each object in the cluster and the cluster centroid C_k. Given a cluster k_i, let the set of objects mapped to that cluster be $\{t_{i1}, t_{i2}, ..., t_{im}\}$. The squared error is defined as

$$seK_i = \sum_{j=i}^{m} \left\| t_{ij} - C_k \right\|^2 \qquad (3.1)$$

Given a set of clusters $K = \{K_1, K_2, ..., K_k\}$, the squared error for K is defined as:

$$seK = \sum_{j=i}^{k} seK_i \qquad (3.2)$$

The outcome of the *k*-means strongly depends on two factors. Firstly, results obtained depend on the first randomly selected centroids. Implication of this being, results obtained on concessive runs on the same data might produce different results even though MacQueen (1967:290) argues that this should not be a huge concern, since only data points on the boundaries of clusters might be influenced. Secondly, different *k*-values will produce different results and the selected *k*-value should correspond to the natural structure of the data, while finding an appropriate *k*-value might be difficult. Too high a *k*-value might lead to over fitting whereas too low a *k*-value might lead to patterns being undetected. The *k*-means algorithm is also sensitive to outliers.

3.2.3 Evaluation of clustering algorithms

No single clustering algorithm can fulfil all clustering requirements, and in view of this, it is important to select a clustering algorithm appropriate for the task. Zaïane *et al.* (2002:28) list the following typical requirements for a good clustering technique:

- Scalability: The cluster method should be applicable to huge databases and performance should decrease linearly with data size increase;
- Versatility: Clustering objects could be of different types – numerical data, boolean data or categorical data. Ideally a clustering method should be suitable for all different types of data objects;

- Ability to discover clusters with different shapes: This is an important requirement for spatial data clustering. Many clustering algorithms can only discover clusters with spherical shapes;
- Minimal input parameter: The method should require a minimum amount of domain knowledge for correct clustering. However, most current clustering algorithms have several key parameters and they are thus not practical for use in real world applications;
- Robust with regard to noise: This is important because noise exists everywhere in practical problems. A good clustering algorithm should be able to perform successfully even in the presence of a great deal of noise;
- Insensitive to the data input order: The clustering method should give consistent results irrespective of the order the data is presented;
- Scale able to high dimensionality: The ability to handle high dimensionality is very challenging but real data sets are often multidimensional.

Mingoti and Lima (2006:1742) conducted an in-depth study in which several clustering algorithms were evaluated. Cluster algorithms were evaluated using criteria such as:

- correct classification per number of variables and clusters with overlapping and without the presence of overlapping;
- average results for correct classification and internal cluster dispersion rates without the presence of cluster overlapping;
- average correct classification rate by number of variables and clusters in the presence of 40% and 60% overlapping respectively;
- correct classification rate for clusters with outliers, with no overlapping present.

In general, they found that the *k*-means algorithm performed better than self-organising maps in their study. In addition, results obtained by the *k*-means method were substantially improved when the initial cluster seeds were selected through the use of a hierarchical clustering method. The *k*-means clustering algorithm portrayed a high level of robustness in situations where the input data set was disconcerted and the occurrence of overlapping also did not seriously influence the results obtained. On the contrary, the SOM networks were negatively impacted in the presence of overlapping. Overlapping occurs when an object belongs to more than one cluster. Both *k*-means and SOM networks were affected when the test data set contained a large number of outliers. Interestingly, it was found that in general, clustering algorithms are more affected by overlapping than by outliers' presence (Mingoti & Lima, 2006:1756).

3.2.4 Evaluation of clustering results

Clustering algorithms process data and generate clusters whether or not natural structure exists within presented data. In the presence of natural structure, some algorithms might be better equipped than other methods to detect underlying patterns within the data, depending on data structure, etc. Appraising clustering results is problematic in view of the fact that clustering data is an undirected data mining technique. Hempel (2002:8) states that with undirected data mining techniques there is no right or wrong answer, and those boundaries are fuzzy. It is the modeller's task to analyse results and determine the significance of the outcome.

According to Berkhin (2002:38), clustering should start by determining the plausibility of clustering the presented heterogeneous data, followed by the selection of appropriate attributes to use during clustering. Results obtained by applying a clustering method can then be assessed by a subject expert. Assessing results consisting of high dimensional data can be very challenging.

Branches of the financial institution in this study will be clustered using the *k*-means clustering algorithm, and the identified clusters of homogeneous branches will be compared in order to identify which branches operate efficiently using DEA. Clustering branches into homogeneous groups assures that when a branch has been identified as operating efficiently among similar branches, the comparison was indeed balanced and not bias towards bigger branches. The following section describes DEA.

3.3 Data envelopment analysis

In the second step of the analytical model, DEA will be used to evaluate the efficiency of clusters of homogeneous branches identified in step one. This section gives a synopsis of DEA, a management science technique that measures the efficiency of a decision making unit (DMU) relative to other units in the population under scrutiny (Cook & Zhu, 2007:692). DMUs may range from different schools in the same district, government departments performing similar operations, or any other units as long as the units under scrutiny perform similar operations and are sensibly comparable to others. In this study, branches of the same financial institution will be measured to determine their operating efficiency.

In the simplest of cases, efficiency may be measured as a ratio of outputs produced to inputs consumed. Baker (2006:157) uses a dairy farm case to illustrate the basic idea of efficiency measure. In this example, there are five dairy farms; every farm has a number of cows that produce milk as illustrated in Table 3.1. Productivity for these farms is measured as the ratio of outputs produced to inputs consumed. Farm 4 obtained the highest productivity ratio with a ratio of 5. Efficiency is merely the normalised form of productivity, and the efficiency of a farm is calculated as the productivity ratio of the farm to the maximum productivity ratio obtained by other farms under investigation. Thus, it is easy to observe that farm 4 has obtained the highest efficiency with a score of 1.00 and farm 3 operates at the lowest efficiency level of 0.70 for the set of farms under investigation.

Decision making unit (DMU)	Cows (Inputs)	Milk (Outputs)	Productivity	Efficiency
Farm 1	15	60	4.00	0.80
Farm 2	10	48	4.80	0.96
Farm 3	20	70	3.50	0.70
Farm 4	12	60	5.00	1.00
Farm 5	16	72	4.50	0.90

Table 3.1 Efficiency measured in the case of a single input/single output model.

The efficiency difference between farm 3 and farm 4 could be the result of different technologies and /or procedures used, or management styles. Thus, it could be said that if farm 3 utilises the exact same techniques as farm 4, farm 3 has the potential to produce 100 litres of milk with 20 cows (20 cows multiplied by the highest attained productivity (5) for all of the farms). Alternatively, farm 3 should use no more that 14 cows to produce the 70 litres of milk, thus saving additional expenses associated with the unnecessary six cows.

Real world problems are much more complex than the milk farm example as companies consume multiple inputs and produce multiple outputs. In the case of multiple inputs, weights are used to combine the inputs into a single number. The same holds for outputs and the ratio in such cases are calculated as the weighted sum of outputs to the weighted sum of inputs. DEA is a valuable tool for measuring efficiency, and in this study DEA will be applied to identify efficient branches.

DEA may be seen as the core of the proposed framework. Results obtained from DEA, labelling a branch as efficient or inefficient, will be used in the third step, the DTI phase, of the proposed framework. Since homogeneous branches are compared to determine the efficiency label for DTI, the same results can be used to benchmark branches in the existing branch network.

3.3.1 The origins of data envelopment analysis

DEA is a management science technique dating back to the mid 1970s. Edwardo Rhodes researched an educational program for disadvantaged students in a U.S. public school as part of his Ph.D. dissertation. The research involved comparing similar schools in terms of outputs achieved, such as increased self esteem in a disadvantaged child, against inputs utilised, akin to time spent by a mother reading to her child. The challenge of estimating the relative efficiency of schools with multiple inputs and multiple outputs in the absence of information on prices has led to the development of the Charnes, Cooper and Rhodes (CCR) DEA model. CCR is the most basic DEA model and is named after Charnes, Cooper and Rhodes respectively. In 1978, DEA in the form of the CCR model was introduced to the world (Charnes *et al.*, 1978). CCR is based on the Farrell (1957) single-output/single-input technical-efficiency measure, converted to the multiple-output/multiple-input case by constructing a single virtual output to a single virtual input relative-efficiency measure (Charnes *et al.*, 1994:4).

3.3.2 Characteristics of data envelopment analysis

DEA offers an alternative approach for extracting information regarding a population. Table 3.2 lists differences between parametric, standard statistical approaches and non-parametric approaches as in the case of DEA (De Leone & Lazzari, 1998:2).

Standard statistical approaches (parametrical approaches)	**Data envelopment analysis (Non-parametric approaches)**
Objective is to optimise a single regression plane through the data.	Optimises on each individual observation with the objective of calculating a discrete piecewise frontier determined by the set of Pareto-efficient DMUs.
Single optimised regression equation is assumed to apply to each DMU.	Optimises the performance measure of each DMU. This results in a revealed understanding about each DMU instead of an average applying to the entire population.
Require the imposition of a specific functional form e.g. a regression equation, a production function, etc. relating independent variables to the dependent variables and require specific assumptions about the distribution of the error terms.	Does not require any assumption about the functional form. Calculates a maximal performance measure for each DMU relative to all the other DMUs in the population with the sole requirement that each DMU lies on or below the external frontier.

Table 3.2 Differences between standard statistical approaches and DEA (De Leone & Lazzari, 1998:2).

Charnes *et al.* (1994:5) list 12 characteristics DEA calculations focus on:

- focus on individual observations in contrast to population averages;
- produce a single aggregate measure for each DMU in terms of its utilisation of input factors to produce desired outputs;
- can simultaneously utilise multiple outputs and multiple inputs with each being stated in different units of measurement;
- can adjust for exogenous variables;
- can incorporate categorical variables;
- are value free and do not require specification or knowledge of a priori weights or prices for the inputs or outputs;
- place no restriction on the functional form of the production relationship;
- can accommodate judgment when desired;
- produce specific estimates for desired changes in inputs and/or outputs for projecting DMUs below the efficient frontier onto the efficient frontier;
- are pareto-optimal;
- focus on revealed best-practice frontiers rather than on central tendency properties of frontiers;
- satisfy strict equity criteria in the relatives evaluation of each DMU.

3.3.3 Data envelopment analysis models

DEA offers the decision maker several advantages. The ability of DEA to highlight best practice units among a population of similar entities, as well as the aptitude to recognise the magnitude of inefficiency associated with inputs and outputs for each entity makes DEA stand out among other analytical techniques (Cooper *et al.*, 2007:14). A high level of flexibility coupled with DEA is another reason for its wide popularity (Samoilenko & Osei-Bryson, 2008:1570). This section describes basic DEA models and various orientations available to the researcher; the model discussion should make the flexibility linked to DEA obvious.

The researcher could use one of several models and orientations the model has to offer. Figure 3.4 depicts different DEA models and orientations available to the researcher. The DEA model selected for application to a problem would depend on the economic assumptions associated with the problem, for example, if the output produced by the DMU is directly equal to the input consumed by the DMU, the researcher would apply a 'constant returns to scale'- model to the problem. In addition to selecting a model, the researcher also has the liberty, for some models, to choose between orientations. Generally, there are three orientations available to choose from, i.e. input-orientated, output-orientated and base-oriented models (Cooper *et al.*, 2007:14). If the emphasis of the study is to minimise input usage while maintaining the same level of outputs, the model is defined as input-orientated. The output-orientated model is the inverse of the input-orientated model and tries to maximise output with given input. Lastly, the base-oriented model deals with input excesses and output deficiencies simultaneously.

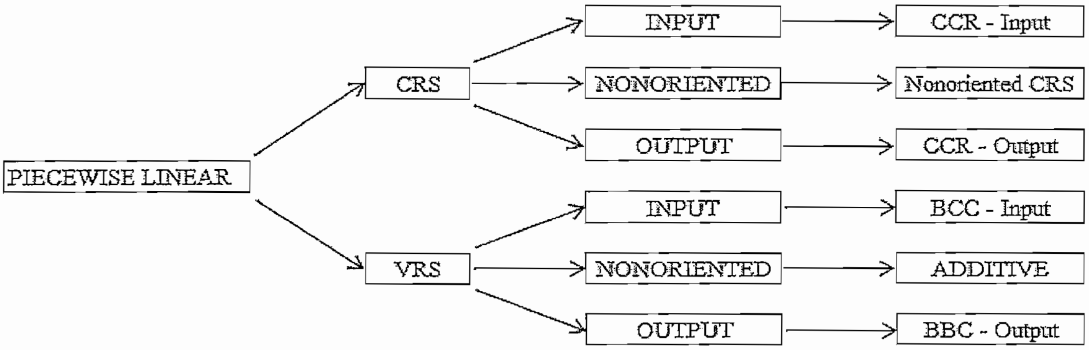


Figure 3.4 Classification and orientation of DEA models (Charnes *et al.*, 1994:66).

Various formulations of DEA models exist allowing alternative approaches to analyse data from an efficiency measurement perspective (Ali & Seiford, 1993:120). The following section introduces the formulation of two basic DEA models, the Charnes, Cooper and Rhodes (CCR) and Banker, Charnes and Cooper (BCC) DEA models. The discussion of both models is based on work done by Cook and Seiford (2009) and Charnes *et al.* (1994:39).

3.3.3.1 The CCR data envelopment analysis model

In the discussion to follow, we assume a set of *n* DMUs to be evaluated, each DMU *j,(j=1,..,n)* using *m* inputs x_{ij} *(i=1,..,m)* and generating *s* outputs y_{rj} *(r=1,..,s).* If the prices or multipliers $\bar{u}_r, \bar{v}_i$ associated with outputs *r* and inputs *i*, respectively, are known, then borrowing from conventional benefit/cost theory, one could express the efficiency $\bar{e}_j$ of DMU *j* as the ratio of weighted outputs to weighted inputs (see equation (3.3)). This is the mathematical form of the benefit/cost theory. The weights $\bar{u}_r, \bar{v}_i$ are used to value the combination of inputs and combination of outputs in (3.3)

$$\sum_r \bar{u}_r y_{rj} \Big/ \sum_i \bar{v}_i x_{ij} \qquad (3.3)$$

The benefit/cost ratio is the basis for the standard engineering ratio of productivity. Important to remember is that in the case of (3.3), multipliers $\bar{u}_r, \bar{v}_i$ are known and can simply be exchanged with actual weight values. Charnes, Cooper and Rhodes developed the CCR DEA model in the absence of known multipliers and proposed deriving appropriate multipliers for a given DMU by solving a particular non-linear programming problem. Specifically, if DMU_0 is under consideration, the CCR model for measuring the technical efficiency of that DMU is given by the solution to the fractional programming problem

$$e_0 = \max \sum_r u_r y_{ro} \Big/ \sum_i v_i x_{io}$$

subject to

$$\begin{aligned} &\sum_r u_r y_{rj} - \sum_i v_i x_{ij} \le 0, && \forall j, \qquad (3.4) \\ &u_r, v_i \ge \varepsilon, && \forall r, i. \end{aligned}$$

where the symbol ε is a non-archimedian value designed to enforce strict positivity on the variables. The above mathematical model is a basic DEA model initially developed by Charnes *et al.* (1978). Because of the assumed constant return to scale, the CCR model is classified as a constant return to scale (CRS) DEA model. The fractional programming model (3.4) may be converted into a linear programming (LP) model as shown in (3.5)

$$e_0 = \max \sum_r \mu_r y_{ro}$$

subject to

$$\begin{aligned} &\sum_i v_i x_{io} = 1 && \\ &\sum_r \mu_r y_{rj} - \sum_i v_i x_{ij} \le 0, && \forall j\,, \qquad\qquad (3.5) \\ &\mu_r, v_i \ge \varepsilon, && \forall r,i. \end{aligned}$$

Performing DEA requires the solution of *n* linear programming problems of the above form, one for every DMU in the population. The graphical representation of the CCR DEA model may be seen in Figure 3.5 and displays seven DMUs in a single input/single output case. The efficient frontier is represented by the solid line passing through DMU #2 and no other point. Only DMU #2 is efficient and is in the reference set of all the other DMUs. The efficiency of DMU #5 can be calculated as AB/AC = 1.8/5 = 0.36, or 36%. Thus, for DMU #5 to move onto the efficiency frontier and be deemed efficient, DMU #5 would have to reduce its inputs by 64% while maintaining the same level of output. A DMU identified as efficient in the CCR input-oriented model, will also be characterised as efficient in the output-oriented CCR model.

DMU	1	2	3	4	5	6	7
X (Input)	2	3	6	9	5	4	10
Y (Output)	2	5	7	8	3	1	7

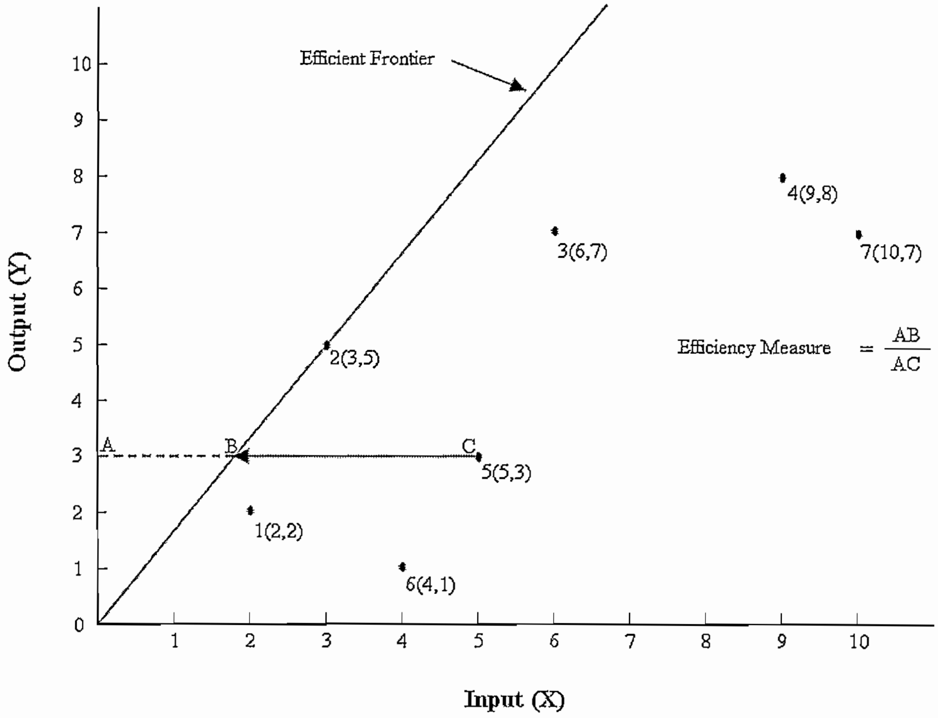


Figure 3.5 CCR model in a single input/single output case (Cook & Seiford, 2009:3).

The next section describes the BBC DEA model that differs from the CCR model in that it takes variable returns to scale into consideration.

3.3.3.2 The BCC data envelopment analysis model

Since the beginning of DEA, various extensions to the CCR model have been proposed of which the BCC model was one. The most apparent reason for extending the CCR model was that CCR was built on the supposition of constant returns to scale. Banker *et al.* (1984) suggested a new model that took variable returns to scale (VRS) into consideration and the model became known as the BCC model, named after its creators. Model (3.6) is the fractional programming representation of the BCC model for a specific DMU_0

$$e_0^{\,*} = \max\left[\sum_r u_r y_{ro} - u_o\right] \Big/ \sum_i v_i x_{io}$$

subject to

$$\sum_r u_r y_{rj} - u_o - \sum_i v_i x_{ij} \leq 0, \quad j = 1,\ldots,n \qquad (3.6)$$

$$u_r \geq \varepsilon, \qquad v_i \geq \varepsilon, \qquad \forall i, r \,,$$

u_o unrestricted in sign.

The linear programming equivalent of (3.6) is shown in (3.7)

$$e^*{}_o = \max \sum_r \mu_r y_{ro} - \mu_o$$

subject to

$$\sum_i v_i x_{io} = 1$$

$$\sum_r \mu_r y_{rj} - \mu_o - \sum_i v_i x_{ij} \leq 0, \quad j = 1,\ldots,n \qquad (3.7)$$

$$\mu_r \geq \varepsilon, \qquad v_i \geq \varepsilon, \qquad \forall i, r,$$

μ_o unrestricted in sign.

Graphically, Figure 3.6 depicts the BBC model for a single input and output. The dotted line passing through DMU #2 is the CCR efficiency frontier. The bold line connecting DMUs 1, 2, 3 and 4 is the BCC efficiency frontier. From this graphical representation, it is easy to see the relation between the CCR and the BCC models. If a DMU is identified as efficient in the CCR model, it will also be characterised as efficient with the BBC model, but the opposite might not always be true (Cooper *et al.*, 2006:85). The BCC model takes variable returns to scale into consideration, and increasing, constant and decreasing returns to scale can be seen in Figure 3.6. The section on the frontier line from DMU#1 up to DMU#2 represents increasing returns to sale, while the section at DMU#2 represents constant returns to scale. DMUs from #2 up towards DMU#4 experience decreasing returns to scale.

DMU	1	2	3	4	5	6	7
X (Input)	2	3	6	9	5	4	10
Y (Output)	2	5	7	8	3	1	7

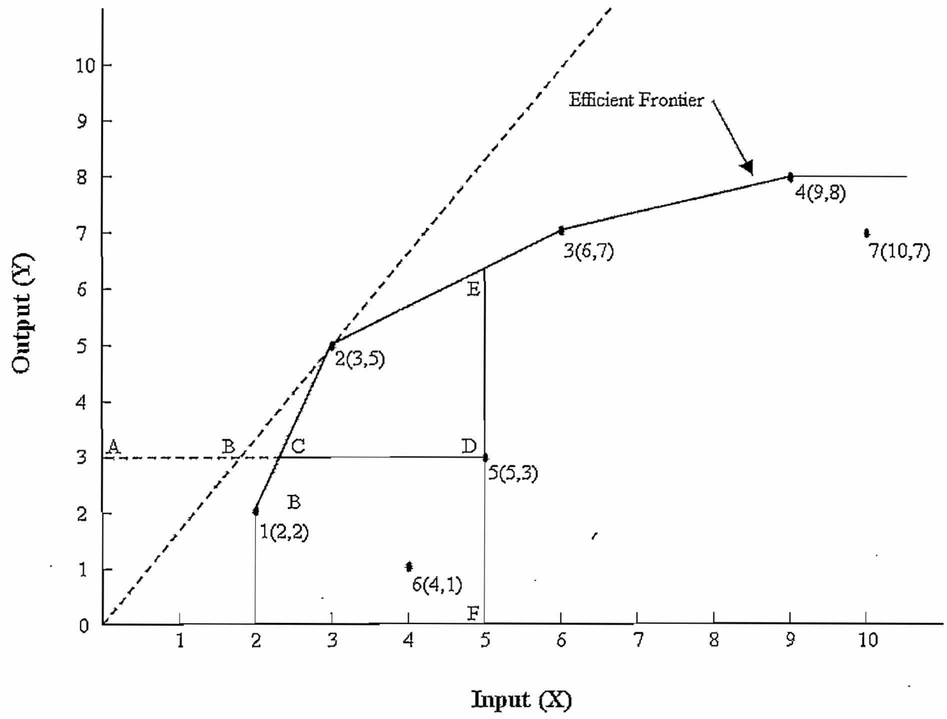


Figure 3.6 BCC model in a single input/single output case (Cook & Seiford, 2008:3).

The CCR and BCC models are the basic DEA models, and with the wide spread success of DEA several extensions to the basic DEA models were rapidly developed. Table 3.3 lists and draws a simple comparison among some of the basic DEA models available.

The benefits of applying DEA as an efficiency measurement tool are generally accepted. However, there are certain limitations associated with the use of DEA the user should be aware of. An example of a limitation is the fact that DEA assumes a certain level of homogeneity among DMUs under investigation. According to Dyson *et al.* (2001:247), this limitation can be overcome by applying CA to the DMUs, as is the case in this study. Dyson *et al.* (2001) report on the limitations of DEA and provide general guidelines as to overcome the deficiencies.

Model	Returns to scale	Envelopment surface	Projection map	Envelopment Metric (Range)	Units invariant	Involves Non-Archimedean
Additive	Variable	Piecewise linear	$Y_o \rightarrow Y_o + S^+$ $X_o \rightarrow X_o - S^-$	L1 $(z \leq 0)$	No	No
Invariant Multiplicativ e	Constant (Log-linear)	Piecewise Cobb-Douglas	$Y_o \rightarrow Y_o e^{-s^+}$ $X_o \rightarrow X_o e^{-s^-}$	e^{L1} $(e^{s+} \geq 1, 0 < e^{s^-} \leq 1)$	Yes	No
Variant multiplicativ e	Constant (Log-linear)	Piecewise Log-linear	$Y_o \rightarrow Y_o e^{s^+}$ $X_o \rightarrow X_o e^{-s^-}$	e^{L1} $(e^{s+} \geq 1, 0 < e^{s^-} \leq 1)$	No	No
BCC Input	Variable	Piecewise Linear	$Y_o \rightarrow Y_o + s^+$ $X_o \rightarrow \theta X_o - s^-$	Radial (Inputs) $(0 < \theta \leq 1)$	Yes	Yes
BCC Output	Variable	Piecewise Linear	$Y_o \rightarrow \phi Y_o + s^+$ $X_o \rightarrow X_o - s^-$	Radial (Outputs) $(\phi \geq 1)$	Yes	Yes
CCR Input	Piecewise Constant	Piecewise Linear	$Y_o \rightarrow Y_o + S^+$ $X_o \rightarrow \theta X_o - s^-$	Radial (Inputs) $(0 < \theta \leq 1)$	Yes	Yes
CCR Output	Piecewise Constant	Piecewise Linear	$Y_o \rightarrow \phi Y_o + S^+$ $X_o \rightarrow X_o - s^-$	Radial (Outputs) $(\phi \geq 1)$	Yes	Yes

Table 3.3 Comparison of basic DEA models (Charnes *et al.*, 1994:45).

Section 3.3 dealt with DEA and the characteristics of the basic DEA models, namely CCR and BCC. In the DEA phase, similar branches are evaluated for efficiency and labelled efficient or inefficient using the CCR DEA model. Measures selected and used during the DEA phase, will be discussed in greater detail in Chapter 5. The efficiency label associated with additional variables will be used during DTI and will be discussed in the following section.

3.4 Predicting branch efficiency using decision trees

The methodology used consists of 3 distinct steps, each with a separate goal and every step equally important. The first step, CA, was important to ensure that similar branches were compared. DEA, the second step, highlighted efficient and inefficient branches. Branch efficiency indicators derived from DEA will be used as input to DTI, the third step. Generating rules from data will be dealt with in this step. The next section of this chapter discusses DTs in greater detail.

3.4.1 Decision tree induction

A DT is a tree structure-representation of a given decision problem and is mainly used to discover underlying data structures, thereafter representing these structures in the form of a tree (Sohn & Moon, 2004:281). Sun and Li (2008:2) explain that non-leaf nodes in the DT structure represent a test on an attribute, while leaf nodes characterise a class. In addition, DTs are also very useful for determining relationships between variables. DTs are built through an iterative process of partitioning data into smaller groups more similar in relation to a certain target variable (Berry & Linoff, 2000:113). The DT algorithm tries to partition the data using every possible split and then selecting the variable that best separates the data. The process of partitioning the data continues until no additional advantage is gained through further data separation.

DT is an easy to understand top-down tree structure where decisions are made at each node. Figure 3.7 depicts a DT consisting of 18 nodes and 5 levels as constructed with SAS Enterprise Miner tree viewer tool. The tree illustrated in Figure 3.7 is a simplified version of a DT classifying whether customers will respond to and place orders from marketing catalogues. The node at the top of the tree is called the root node, while nodes at the bottom of the tree present the final classification (leaf-nodes) and may be either categorical, or continuous. Nodes between the root and leaf nodes, called "non-leaf" nodes, present a test on an attribute and determine the path a certain case will follow from the root node to the leaf nodes. Leaf-nodes represent the class the case belongs to, in the case of Figure 3.7, whether a customer will respond to a marketing campaign or not. After the DT has been constructed and presented in a visual form, the tree structure may be converted into a straight forward set of decision rules (Samoilenko & Osei-Bryson, 2008:1572).

Trees are classified as either a classification, or a regression tree. In cases where the target variable is a discrete variable, for example, grading customers as good or bad according to their credit rating, the tree is classified as a classification tree. Thus, a classification tree has one of several discrete values as likely outcomes. Trees with a numerical, continuous target variable, for example, the amount a customer is likely to spend at a sale, are classified as a regression tree (Seol *et al.*, 2007:434).

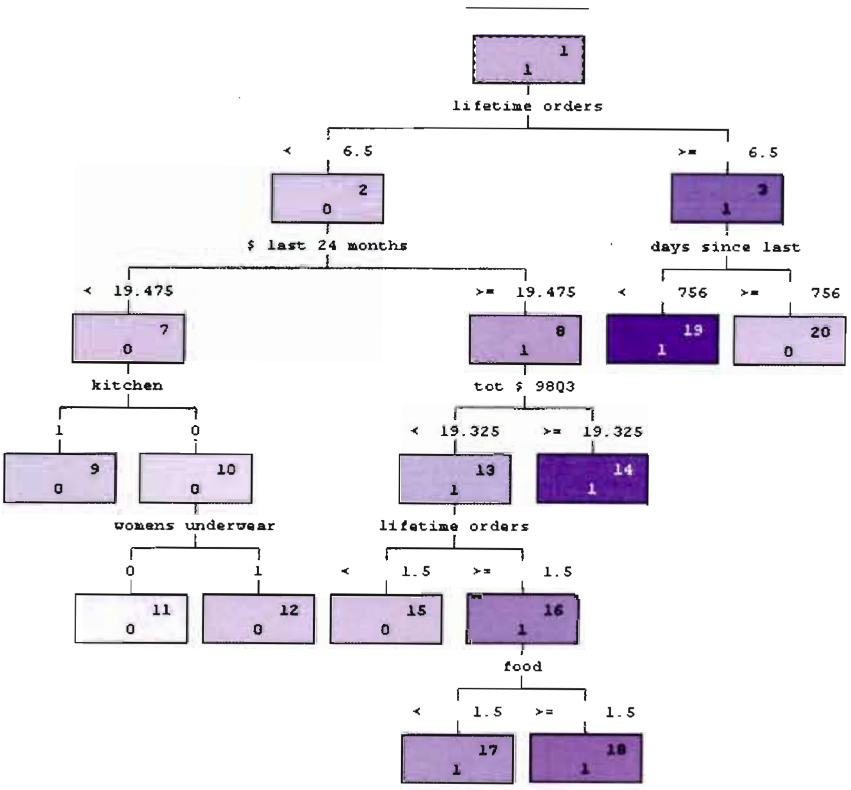


Figure 3.7 Graphical display of a DT classifying whether a customer will respond to a marketing campaign (Data Miners, Inc., 2004).

Classification trees attempt to enhance the purity of a categorical variable within the child node. In the case of a regression tree, the goal is to decrease the variance in the target values of the child nodes (Berry & Linoff, 2004:170). This research will make use of a classification tree, as it will classify whether a branch operates efficiently or not. The rest of this chapter will therefore focus on classification trees.

DT induction is categorised as a supervised data mining technique in contrast to unsupervised data mining techniques like CA, where the predicted value is not known beforehand. Training data, i.e. data where the value of the target variable is known beforehand, is used to train the DT (Sun & Li, 2008:2). The DT is then constructed in two phases. During the first phase of DTI, training data is used to train the model and a function is derived that maps a case in the data to the target variable. Once the DT is constructed, test data is used to assess the accuracy of the derived model. If the accuracy of the model is of an acceptable level, the DT can be applied to unseen data. This comprises the second step of the process. Han and Kamber (2006:293) provide a basic algorithm that may be used to construct a DT. The algorithm presented in Figure 3.8 requires three parameters.

```
create a node $N$
if tuples in $D$ are all of the same class, C then
        return $N$ as a leaf node labelled with the class C;
if attribute_list is empty then
        return $N$ as a leaf node labelled with the majority class in $D$; // majority voting
apply attribute_selection_method ($D$, attribute_list) to find the "best" splitting_criterion;
label node $N$ with splitting_criterion;
if splitting attribute is discrete-valued and
        multi-way splits allowed then // not restricted to binary trees
        attribute_list ← attribute_list – splitting_attribute; // remove splitting_attribute
for each outcome $j$ of splitting_criterion
        // partition the tuples and grow sub trees for each partition
        let $Dj$ be the set of data tuples in $D$ satisfying outcome $j$; // a partition
        if $Dj$ is empty then
                attach a leaf labelled with the majority class in $D$ to node $N$;
        else attach the node returned by Generate_decision_tree($Dj$, attribute_list) to node
Endfor
return $N$;
```

Figure 3.8 A basic algorithm for creating DT (Han & Kamber, 2006:293).

The first parameter, *D*, is a data set containing cases with training data and their associated class labels. Secondly, the *attribute_list* parameter is required. This parameter describes the cases presented in the data set. The attributes in the *attribute_list* are candidate attributes to be considered throughout the data partitioning process. The final parameter and the foundation of any DTI algorithm is the *attribute_selection_method*. This parameter dictates the partitioning phase of the induction process and ultimately the final outcome. During the partitioning phase, the *attribute_selection_method* determines which attribute from the *attribute_list* will be used as the *splitting_attribute*. The *attribute_selection_method* may be a heuristic, such as information gain, Gini index or gain ratio.

Among data mining techniques, DT is almost certainly the most popular mining method (Chen *et al.*, 2003:202). The reputation of this method is due largely to the fact that DTI offers the analyst several advantages over other mining techniques, thus guaranteeing DTI methods to be employed by just about any data miner. DTs are usually the first tool analysts use during data mining expeditions, since DTs perform exceptionally well at highlighting important variables in a data set. However, other methods might be called upon during a later phase of a mining expedition. The first step of exploring the data is usually conducted with the help of a DT (Berry & Linoff, 2004:165). DTs also do not require prior knowledge about the nature of the data set (Zhao & Zhang, 2008:1956).

DTs are renowned for being well organised, easily interpretable by data analysts, computationally inexpensive and competent in the presence of noisy data (Chen *et al.*, 2008:4839). The manner in which DTs evaluate numerical attributes, make them immune to differences of scale and outliers usually associated with numerical attributes. The additional benefit of this characteristic is less time required during the data preparation phase. On the other hand, data mining methods like neural networks and CA are very sensitive to numerical attributes and depend a great deal on data preparation (Berry & Linoff, 2000:120).

3.4.2 Basic decision tree algorithms

Many DT algorithms have been proposed, the most popular methods for DTI being ID3, C4.5 and CART (Walsh, 2005:2-42). These algorithms create DTs in a top-down recursive divide and conquer manner. Iterative Dichotomiser (ID3) was developed by Quinlan in the early 1980s. Quinlan also presented the algorithm C4.5 as an improvement on his earlier work ID3. C4.5 uses gain-ratio for selecting splitting attributes, as opposed to information-gain, used in the ID3 DT algorithm. C4.5, a well known DT algorithm, uses a post-pruning method to improve the accuracy of trees (Walsh, 2005:2-42). During the same time, another algorithm, CART (Classification and Regression Trees), was presented by Breiman, Friedman, Olshen and Stone (Han & Kamber, 2006:243). The CART algorithm applied the Gini index measure to calculate the impurity of a data set. Although these algorithms were developed independently, they exercise similar logic for constructing DTs.

3.4.3 Attribute selection measures

DTs are grown by dividing data into smaller groups that are homogeneous in respect to a target variable. Dividing the data set into smaller partitions, entails a process of selecting an attribute that does the best job of partitioning the data; this process of selecting the attribute that would best partition the data is called the attribute selection measure. The method selected to identify the splitting attribute has the biggest impact on the creation of the DT. It ultimately dictates the final structure of the tree and in addition, also determines aspects such as number of branches grown at a splitting node.

For the remainder of this section, the following notation will be used: Let D be a set of training cases with a target variable, the class, the training cases belong to. Suppose the class label attribute has m distinct values defining m distinct classes, C_i (for $i = 1,\ldots,m$). Let $C_{i,D}$ be the set of cases of class C_i in D. Let $|D|$ and $|C_{i,D}|$ denote the number of cases in D and $C_{i,D}$ respectively.

3.4.3.1 Information gain

Information gain, rooted in information theory, is a measure of how disorganised a system is and is known as entropy reduction (Han & Kamber, 2006:297). Mathematically this may be expressed as

$$Info(D) = -\sum_{i=1}^{m} p_i \log_2(p_i) \qquad (3.7)$$

where p_i is the probability that a randomly selected tuple from D belongs to class C_i and is calculated as $|C_{i,D}|/|D|$. $Info(D)$ is therefore based on the proportion of tuples in D belonging to class C_i and is known as the entropy of D. Splitting D would require a test on all attributes in D to determine how much additional information would be required if the attribute was selected as the splitting attribute. Assume attribute A, a categorical attribute from D, was selected to perform the split on. A, having v distinct values $\{a_1,\ldots,a_v\}$ would then split D into $\{D_1,\ldots,D_v\}$, where D_j contains those tuples in D that belong to a_j of A. $Info_A(D)$ in equation (3.8) may be used to measure the additional information post partitioning required to accurately classify a tuple.

$$Info_A(D) = \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times Info(D_j) \qquad (3.8)$$

Information gain is the difference between pre-split information requirement and post-split information requirement and is given in (3.9). It shows how much information was required to accurately classify a tuple before splitting D and after splitting D.

$$Gain(A) = Info(D) - Info_A(D) \qquad (3.9)$$

The attribute with the highest information gain is chosen as the splitting attribute for that node (Han & Kamber, 2006:300). Suppose attribute A was a continuous variable in data set D instead of a categorical variable as just explained, then a splitting point opposed to a splitting attribute has to be determined. In such cases, the midpoint between neighbouring numerical values has to be considered as a possible *split_point*. Consequently D will be split into two partitions, the first partition where $A \leq$ *split_point* and the second partition where the numerical values of $A >$ *split_point*. Equation (3.10) is used to calculate splitting points in the numerical attribute.

$$\frac{a_i + a_{i+1}}{2} \qquad (3.10)$$

Every splitting point in A has to be evaluated with equation (3.8), and since the splitting point partitions the data into two partitions v, equation (3.8) will have two values, one partition containing values smaller than, or equal to the splitting point and a second partition containing values greater than the splitting point.

Information gain was used in the ID3 algorithm, but has a major drawback. When an attribute selected as the splitting attribute were a discrete valued attribute, a branch would be grown for every distinct value within the splitting attribute. The effect of this was that when the attribute consisted of many distinct discrete values, the algorithm was biased towards splitting on these attributes. Splitting on an attribute with many distinct values resulted in a split with many branches consisting of a single tuple, thus resulting in a 100% pure branch. Information gained on a pure branch would be maximal. Gain ratio, the successor to information gain, does not have the same deficiency.

3.4.3.2 Gain ratio

The C4.5 DT algorithm utilised gain ratio as an attribute selection measure and was an improvement on the information gain attribute selection measure (Han & Kamber, 2006:301). This algorithm strived to overcome the subjective drawback associated with the information gain algorithm. $SplitInfo_A(D)$, equation (3.11), yields the information gain when splitting the data on attribute A, the difference being that the split value of A is calculated with respect to all the tuples in D, whereas information gain was calculated based on the same partition.

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times \log_2\left(\frac{|D_j|}{|D|}\right) \tag{3.11}$$

$$GainRatio(A) = \frac{Gain(A)}{SplitInfo(A)} \tag{3.12}$$

The attribute with the highest gain ratio is selected as the splitting attribute. Gain ratio was an improvement on information gain and was adjusted to compensate for the biasness towards multi valued attributes associated with information gain. A weakness identified with gain ratio is its preference to create partitions where one partition was much smaller than other partitions (Han & Kamber, 2006:303).

3.4.3.3 Gini index

The CART DT algorithm makes use of the Gini index that generates a DT by performing binary splits on attributes (Han & Kamber, 2006:303). Gini index employs equation (3.13) to compute the impurity of a data set, where p_i is the probability that a case in data set D belongs to class C_i and is calculated as $|C_{i,D}|/|D|$.

$$Gini(D) = 1 - \sum_{i=1}^{m} p_i^{\,2} \tag{3.13}$$

The Gini index creates binary splits on attributes, thus for a discrete valued attribute A, comprising of v distinct values $\{a_1, ..., a_v\}$, every combination must be assessed as a splitting possibility. Han and Kamber (2006:303) state that there are $2^v - 2$ possible combinations to be evaluated, for example, attribute A, describing a customer's health having three distinct values, namely {good, average, poor}, have the following six possible splits:

- good, average
- good, bad

- average, bad
- good
- average
- bad

$$Gini_A(D)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \qquad (3.14)$$

For every attribute, every possible split of data set D must be evaluated. Binary partitioning D would result in two partitions, D_1 and D_2, for example, $D_1 = health \in \{good, average\}$ and D_2 would contain all elements of health that are not in D_1. Equation (3.13) is used to measure the impurity of the potential new partitions by calculating the weighted sum of impurity for every partition. The partition yielding the lowest Gini index is selected as the splitting subset.

Continuous attributes are handled in the same manner as with information gain. A *split_point* is calculated on neighbouring values, where $A \leq$ *split_point* would be allocated to D_1 and $A >$ *split_point* would be allocated to D_2. Equation (3.14) is used to determine purity of such a split. Reduction in impurity is calculated using equation (3.15), and the attribute that maximises the reduction in impurity is selected as the splitting criteria.

$$\Delta Gini(A)=Gini(D)-Gini_A(D) \qquad (3.15)$$

As with attribute selection measures previously mentioned, the Gini index also has a drawback associated with its use. The Gini index has difficulty when the number of classes is large and has a tendency to be biased towards multi valued attributes (Han & Kamber, 2006:304). Although the three methods described have some disadvantages linked to them, they all produce fairly good results.

3.4.4 Tree pruning

DTs are grown using training data where the class labels of the target variable are known beforehand. The grown tree constructed from the training data could embody irregularities found within the training data. When a DT bears some sort of irregularity on account of the training data, it is said that the tree has been over fit to the training data and might not be accurate on new data. To overcome the problem of over fitting, trained trees undergo a process known as pruning.

There are two methods of tree pruning, namely pre-pruning, also known as bonsai techniques, and post-pruning (Berry & Linoff, 2000:118). The pre-pruning approach proactively restricts the growing process of the tree by performing a test on the node before continuing with partitioning. Tests determining termination of the growing process may be as simple as a minimum node size count, but whatever the computation, the result should be higher than a pre-specified threshold. If the result did not satisfy the threshold, the growth of the tree is stopped and the node is labelled a leaf.

Post-pruning methods on the other hand, allow the full tree to be grown, and then prune branches not providing additional predictive power to the tree. The CART DT algorithm utilises a form of post-pruning when growing DTs (Berry & Linoff, 2004:185). CART depends on a measure called adjusted error rate to identify weak branches of a node and mark them for pruning. The resulting pruned tree is kept and the process is repeated until the tree is pruned all the way to the root node. The sub-trees that have been created are then evaluated using the validation data set, and the tree performing best is selected as the winning tree.

The C5 DT algorithm also applies post-pruning during DTI. C5 unlike CART, does not make use of a validation data set to prune the grown tree, but instead use the same data used during construction for pruning the over-fit tree (Berry & Linoff, 2004:191).

DTs are versatile tools that may be applied to almost any mining problem. In this methodology, DTs will be utilised to link internal and external data and in doing so, generate easily understandable rules that can be interpreted by management.

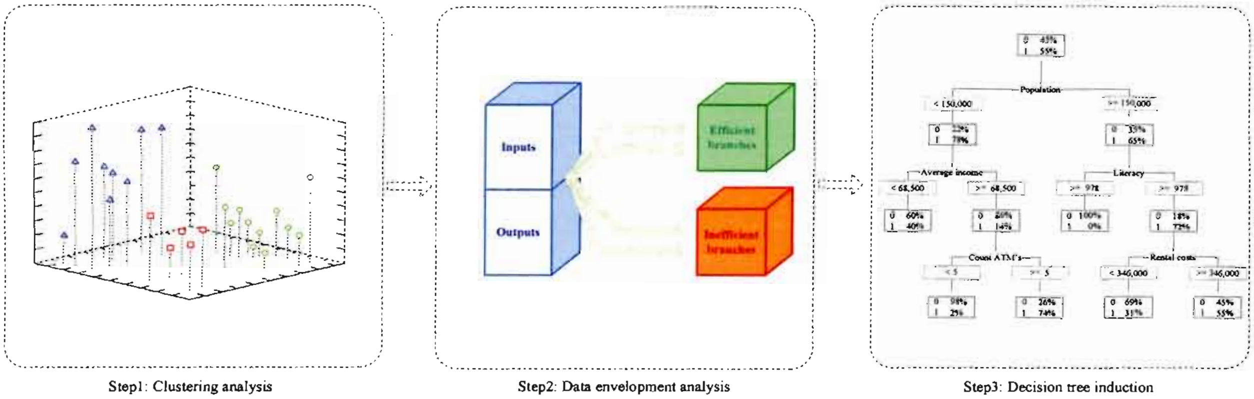
3.5 Chapter conclusion

The literature study conducted in Chapter 2 identified CA, DEA and DTI as techniques that when used in a cohesive manner, can create a powerful analysis tool. It was decided that a combination of these three methods will be applied to aid in the branch management process. Chapter 3 investigated CA and evaluated several clustering methods. Section 3.3 discussed advantages of using DEA. This examination made it clear that data envelopment offers several advantages over other comparison methods, such as regression analysis, and that additional insight into the branches network may be obtained through the use of DEA. The chapter ended with a discussion on DTI. This final step in the methodology will combine insights gained from internal data, with external data describing the market environment for generating rules in aid of the branch expansion process. The next chapter will discuss the research design, attributes selected during each step of the methodology and data manipulation done during the research.

Chapter 4 - Identifying homogeneous branches

4.1 Introduction

Chapters 4, 5 and 6 will be dedicated to the empirical study performed and each chapter will cover in detail a step of the analytical model (Figure 4.1). Figure 4.1 divides the model logically into three steps, starting with clustering analysis (CA) in step 1, identifying efficient branches in step 2 and decision tree induction (DTI) in step 3. This chapter reports on the empirical research relating to CA. CA is the first analytical method used, and the objective is to identify homogenous branches within the financial institution.



Step1: Clustering analysis

Step2: Data envelopment analysis

Step3: Decision tree induction

Figure 4.1 Graphical illustration of analytical model to aid financial institutions in managing the branch network.

The chapter starts with a motivation for the need to identify homogeneous branches and highlights deficiencies of the current segmentation used at the financial institution. Such drawbacks necessitate the application of CA in order to identify homogeneous branches.

Determining which branches of the financial institution operate efficiently, is the mainstay of the analytical model and will be accomplished through the use of data envelopment analysis (DEA), step 2 in Figure 4.1. DEA relies on the assumption that decision making units, branches of the financial institution, when compared, share a high level of similarity (Samoilenko & Osei-Bryson, 2008:1568). Since branches are situated in various different locations, market environments, a logical assumption is that the customer base served by different branches may vary. Consequently, branches differ in the types of services they offer and the customer base they serve. Thus, to improve the discriminatory power of DEA, the branches of the financial institution will be clustered in an attempt to identify similar branches.

CA is applied to the branch information prior to the application of DEA, thereby ensuring that the branches compared are homogeneous. Figure 4.1 displays the analytical model with CA (step 1) a pre-process to DEA. In Chapter 2, a literature study was performed to identify alternative methods for selecting homogeneous branches, and CA was identified as a good method for grouping similar objects. In this study, similar branches of a financial institution will be grouped together. Chapter 3 examined alternative clustering techniques, and *k*-means clustering was selected as method for creating groups of homogeneous branches.

The financial institution operates close to four hundred branches across South Africa. While some of the branches operate in urbanised areas such as Sandton, which is at the heart beat of the South African economy, other branches operate in rural towns, for instance Ventersdorp, within a much smaller society. Therefore, it is of the essence that when comparing branches, they should be similar by nature.

A generic branch segmentation classifying branches by their overall characteristics, already exists within the financial institution. These characteristics include number of transactions, rand value of transactions, number of sales, etc. It was evident that there was scope for improving the current segmentation to improve the grouping of similar branches.

Firstly, it was identified that segmentation should take into account customers that transact at a branch. A valid assumption is that branches with a larger customer base will have more profit potential. A customer is allocated to a branch as a transacting customer where he/she performs most transactions. This data can therefore be seen as part of the active customer base per branch.

Secondly, and possibly just as important, is the number of accounts housed per branch. Since a single customer may have more than one account of various types, it was important to include more account characteristics to the segmentation model. By adding account characteristics to the model, more detail was added, thus allowing for better segmentation.

Finally, at this financial institution, segmentation is done by analysing measures of branch activity. Branch activity is measured by looking at variables such as number of transactions, number of accounts acquired, staffs complement and rand value of transactions. As this classification is a good indication of actual branch activity, it will be included in the new segmentation model.

Section 4.2 introduces the data available for CA, followed by Section 4.3 describing the empirical study at the hand of the SEMMA methodology. The chapter ends with a discussion on the results and insights gained from CA.

4.2 The empirical study methodology applied during this research

This study follows a methodology proposed by the SAS Institute. The Institute is an industry leader in data mining and analytical solutions and utilises the methodology known as SEMMA. The acronym refers to a five step process described by Fernandez (2003:10) as follows:

- Sample the data by extracting a portion of a data set large enough to contain the significant information but small enough to allow for easy manipulation of data;
- Explore the data by searching for unanticipated trends and anomalies in order to gain understanding and ideas;
- Modify the data by creating, selecting and transforming the variables to focus the model selection process;
- Model the data by allowing the software to search automatically for a combination of data that reliably predicts a desired outcome;
- Assess the data by evaluating the usefulness and reliability of the finding from the data mining process.

The SEMMA model will be applied throughout the empirical study of this research. Data sampling, exploring and modifying are collectively known as data preparation and usually require up to 60% of the total research effort (Cabena *et al.*, 1998:47). According to Pyle (1999:3), the quality of the models depends mostly on the content of the data. The SAS SEMMA methodology is regarded most suited for this purpose. The following section will give details about the CA conducted as part of this research.

4.3 Data available to identify similar branches

Data for the research project was provided by the financial institution the research is based on. Information from 398 retail branches across South Africa for the period 1 January 2007 to 31 December 2007 was used.

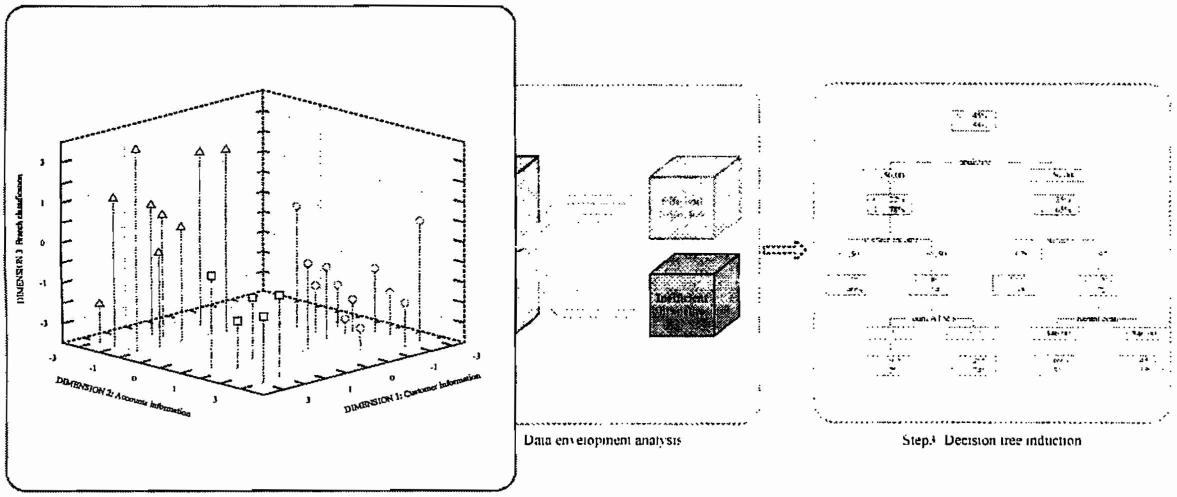
Data obtained for CA was a trade-off between data availability and data requirements. Table 4.1 lists the variables originally considered for CA and gives a brief description of the variables. Other variables that would have made for interesting analysis include branch characteristics such as floor size, number of staff members working at the branch, number of various queries handled at the enquiries desk, etc. The original dataset considered for clustering similar branches contained fourteen variables describing the 398 retail branches of the financial institution. Branch name (br_name), province the branch operates in (br_province) and name of town the branch operates in (br_town), although listed in Table 4.1, are merely descriptive and will not be used for clustering purposes.

Relevance to clustering analysis	**Variable name**	**Description**
Uniquely identifies branch	br_num	Branch number
Branch name	br_name	Branch name
Province branch operates in	br_province	Province branch is located in
Name of town the branch operates in	br_town	Town branch is located in
Description of surrounding area branch operates in (Financial institution criteria used)	br_area_type	Branch area classification
Measure of branch activity. Generic branch segmentation currently used within the financial institution to classify branches in relation to number of transactions, rand value of transactions, number of sales, etc.	br_classification	Branch type classification
Number of accounts housed at the branch split according to type of account. Variables describing a branch from an account dimension.	br_freqCA	Number of current accounts at branch
	br_freqSA	Number of savings accounts at branch
	br_freqHL	Number of home loan accounts at branch
	br_freqINV	Number of investment accounts at branch
	br_freqABF	Number of asset based accounts at branch
	br_freqCAMS	Number of credit card accounts at branch
	br_freqMM	Number of money market accounts at branch
The active customer base transacting at the branch, split according to customer profile. Variables describing a branch from a customer perspective.	br_primary_CG1	High value customers also referred to as affluent customers and have private bankers assigned to them.
	br_primary_CG2	High value customers also referred to as middle market customers. Do not have private bankers.
	br_primary_Markets	Individuals also referred to as mass market customers and do not fit into the affluent or middle market customer bracket.
	br_primary_SmallBus	Small and medium businesses

Table 4.1 Data available for CA.

This chapter started with three arguments branch clustering will be based on. These arguments can logically be seen as three overarching dimensions for clustering, and step 1 in Figure 4.2 graphically illustrates the three dimensions. Every dimension, comprising of several attributes, describes a branch from a different perspective.

Dimension 1, the customer dimension, contains information describing the branch from a customer's point of view and is made up of 4 attributes, namely number of affluent customers (br_primary_CG1), number of high valued customers (br_primary_CG2), number of average customers (br_primary_Market) and number of customers classified as small businesses (br_primary_SmallBus). It should be noted that these customer-related variables are the number of customers actively generating transactions at the branch, thus transacting customers per branch. A customer is allocated to a branch where he/she performs most transactions, and may therefore be seen as part of the active customer base of that branch. When a customer does not use a branch for transacting, he/she will be allocated to the branch where the relevant account was opened.


DIMENSION 3: Branch classification
3
1
0
-1
-3
-3
-1
0
1
3
DIMENSION 2: Accounts Information
3
1
0
-1
-3
DIMENSION 1: Customer Information
Inefficient
Data envelopment analysis
Step3 Decision tree induction

Step1: Clustering analysis

Figure 4.2 Clustering branches according to three dimensions.

The assertion is made that branches with a larger customer base will have more profit potential. Accommodating this in the clustering process, branches with a similar customer base will be clustered together. Variables in this group will be a good indication of how profitable a branch can be.

The second dimension used to describe a branch is the account dimension. This dimension contains information about the number of accounts housed at a branch and is made up of 7 attributes, namely number of current accounts (br_freqCA), number of savings accounts (br_freqSA), number of home loans (br_freqHL), number of investment accounts (br_freqINV), asset based finance accounts (br_freqABF), number of credit card accounts (br_freqCAMS) and the number of money market accounts (br_freqMM). A description of these variables is given in table4.1.

Since a single customer may have various types of accounts, it was important to include both account characteristics, number of accounts and account types, in the clustering model. Different account types provide for different transactional behaviour and therefore different profit potential. By adding account characteristics to the model, more detail is added allowing for better segmentation.

Dimension three consists of one variable, the branch classification (br_classification). Branch classification is a classification internal to the financial institution. A branch may have one of the following classifications:

- Classic branches: the majority of branches are classified as classic branches delivering a wide range of services to mass market customers;
- Express branches: a slimmed down version of the classic branch providing specially selected services;
- Flagship branches: the pride of the financial institution delivering all services across all customer segments;
- Premier branches: smaller version of the flagship branches and mostly located in prestige market environments such as shopping malls, etc;
- Branch in a box: mobile operated branches in an attempt to deliver banking services to the previously disadvantaged;
- Mobile branches: similar to "branch in a box" branches but deliver additional services;
- Campus branches: situated at a university campus and focus on the new generation;
- In store branches: located inside stores and therefore share floor space with other retailers. There are only six of these branches.

Depending on the services delivered by a branch and branch characteristics such as number of transactions, number of accounts acquired, staff compliment and rand value of transactions, a branch is classified into one of the above categories. As this branch classification is a good indication of actual branch activity, it will be included in the model.

The following sections describe the CA done in this study at the hand of the SAS SEMMA data mining methodology.

4.3.1 Sampling the branch clustering data

In general, sampling is performed to extract a small sample of data from the data population that will still represent the characteristics of the population. This is done as working with a smaller data set is easier. In this study though, as the branch information data set only contains information pertaining to 398 branches, which is a relatively small data set, it was decided that the full data set will be used for the remainder of the CA process.

4.3.2 Exploring the branch data

The goal of data exploration is to gain insight into the data as well as identifying important variables. Table 4.2 lists the results of the basic data exploration done. Thirteen variables, two of which are character and eleven are numeric will be considered for CA. A final set of variables will be presented at a later stage.

CA cannot deal with categorical values therefore, if the character variables are deemed important for inclusion in CA, these variables will have to be converted into a numerical equivalent (Pyle 1999:75). Modification of these variables will receive attention during the modification phase of CA and is in relation to the SEMMA model described in Section 4.2.

Another observation from the basic data exploration is the fact that branch area type (br_area_type) and branch classification (br_classification) variables contain missing values that would require attention during the modification phase, if these variables were to be used for the remainder of the CA.

Variable	N	Mean	Std Dev	Min	Max	Distinct	Empty	Type
Area type branch operates in (br_area_type)	398	n / a	n / a	n / a	n / a	4	1	C
Branch classification (br_classification)	398	n / a	n / a	n / a	n / a	8	8	C
Number of current accounts (br_FreqCA)	398	1,712	1,365	-	7,616	385	0	N
Number of savings accounts (br_FreqSA)	398	3,241	3,345	-	24,987	384	0	N
Number of home loans accounts (br_FreqHL)	398	1,411	1,431	-	11,594	379	0	N
Number of investment accounts (br_FreqINV)	398	814	747	-	6,478	352	0	N
Number of asset based finance accounts (br_FreqABF)	398	227	201	-	1,199	276	0	N
Number of credit card accounts (br_FreqCAMS)	398	2,037	1,646	-	8,533	379	0	N
Number of money market accounts (br_FreqMM)	398	1	2	-	18	14	0	N
Number of affluent customers (br_Primary_CG1)	398	18	36	-	326	84	0	N
Number of high value. customers (br_Primary_CG2)	398	49	63	-	342	130	0	N
Number of average customers (br_Primary_Markets)	398	988	1,015	-	6,409	367	0	N
Number of small business customers (br_Primary_SmallBus)	398	98	96	-	534	206	0	N

Table 4.2 Initial data profiling of branch data used for clustering.

In Table 4.2 it can be seen that in general, branches have a perceptible amount of accounts more than customers. It is understandable, as one customer typically has more than one account. Since clustering is a distance based function, this feature of the data will have a stern impact on the results as variables analysed are measured in different units and over different ranges. For example, looking at the mean column in Table 4.2, customer descriptive columns have relatively smaller mean values compared to mean values describing number of accounts. In numerical terms, this is an indication that variables describing accounts, since most of them have a bigger numerical value than the number of customers, are more important than the customer related variables. This observation is important and Berry and Linoff (2004:364) propose scaling the variables for consistency in order to resolve this problem. Scaling of variable values will receive attention during the modification phase of CA.

4.3.2.1 Correlation analysis in order to remove redundant variables

The relationship among variables was investigated during the second phase of data exploration. Correlation measures how values of one variable change as values of another variable change and is expressed as a number ranging from -1 to +1, where a correlation of ± 1 indicates perfect predictability (Pyle, 1999:317). The relationship analysis of variables is done to remove redundant variables (Han & Kamber, 2006:289). A variable becomes redundant when two variables share a high level of similarity, therefore any one of these may be removed from further analysis. The mathematical expression of the calculation of correlation is shown in equation 4.1, assuming a data set containing 2 variables, (x_i,y_i), $i = 1, 2, \ldots, n$.

$$r = \frac{\sum xy - \frac{1}{n}\left(\sum x \sum y\right)}{\sqrt{\sum x^2 - \frac{1}{n}\left(\sum x\right)^2}\sqrt{\sum y^2 - \frac{1}{n}\left(\sum y\right)^2}} \qquad (4.1)$$

A positive linear correlation of 1 indicates that the two variables are completely linearly related, furthermore, that with an increase in variable x, variable y will also increase, although not necessarily by the same amount. A linear relationship of -1 indicates a perfectly predictable relationship, but the values of the variables move in opposite directions. In general, a correlation value of between +0.3 and -0.3 indicates a weak relationship, and only when the correlation is greater than +0.8 or less than -0.8, does the value indicate a good fit (Pyle 1999:317). The square of the correlation, denoted by the symbol r^2, is a more useful measure expressing the amount of explanatory power one variable has about the value of another.

The squared correlation coefficient is listed in Table 4.3. From this analysis it can be seen that variable *number of savings accounts* (br_freqSA) and variable *number of mass market customers* (br_primary_Markets) have a high correlation value with $r^2 \approx 0.93$ or 93%. A threshold value of $r^2 = 80\%$ was used, and when two variables have a correlation higher than the set threshold, one of the variables will be removed from further analysis.

Variable name	br_freqCA	br_freqSA	br_freqHL	br_freqINV	br_freqABF	br_freqCAMS	br_freqMM	br_primary_CG1	br_primary_CG2	br_primary_Markets	br_primary_SmallBus
br_freqCA	1										
br_freqSA	0.36	1									
br_freqHL	0.47	0.76	1								
br_freqINV	0.57	0.74	0.59	1							
br_freqABF	0.73	0.15	0.30	0.30	1						
br_freqCAMS	0.74	0.32	0.57	0.44	0.63	1					
br_freqMM	0.32	0.00	0.01	0.09	0.24	0.15	1				
br_primary_CG1	0.27	0.00	0.00	0.04	0.18	0.11	0.62	1			
br_primary_CG2	0.54	0.02	0.05	0.14	0.65	0.38	0.46	0.49	1		
br_primary_Markets	0.30	0.93	0.77	0.65	0.12	0.33	0.00	0.01	0.01	1	
br_primary_SmallBus	0.72	0.07	0.15	0.26	0.73	0.49	0.47	0.41	0.77	0.04	1

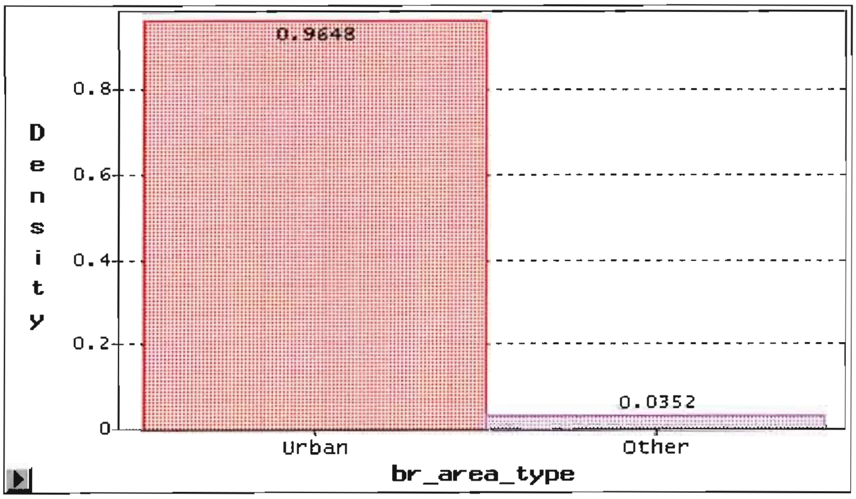
Table 4.3 Correlation analysis of 11 numeric variables considered for CA.

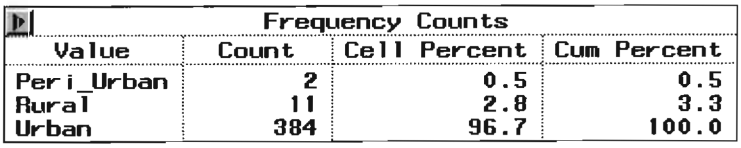
A possible explanation for this occurrence might be the fact that an average customer's surplus funds are usually not adequate for alternative investments, which usually is pricier, and therefore are placed into a savings account. Having funds in a savings account also allows for easier access if required by the customer. In addition, affluent, wealthier customers in customer segments (br_primary_CG1 and br_primary_CG2) usually opt for alternative investments which are more expensive but in the long term yield better returns.

Given the *number of savings accounts* (br_freqSA) and the *number of mass market customers* (br_primary_Markets) have a strong relationship, the *number of market customers* was identified as being redundant and will therefore be excluded from further analysis. The objective is to create clusters with a high level of similarity. It was therefore decided to keep the *number of savings account* variable, as it correlated better with the others, thus resulting in better quality clusters. All other variables less than the threshold value will be included for the remainder of the analysis.

4.3.2.2 Removing variables that do not contain information

Variables not containing information that aid in the process of discriminating between records should, according to Berry and Linoff (2000:135), be ignored for data mining purposes. A histogram shows how often a value, or range of values, occurs for a given variable. During this phase of data exploration, a histogram was generated for every variable in Table 4.2, and it was discovered that 96.7% of the values from the variable describing the area a branch operates in (br_area_type), had the same value. Figure 4.3 shows the histogram and frequency counts for this variable. The br_area_type is an area code internal to the financial institution, and since all the branch information received relates to retail branches, this result was expected. It was decided to remove the br_area_type from further analysis.





Frequency Counts

Value	Count	Cell Percent	Cum Percent
Peri_Urban	2	0.5	0.5
Rural	11	2.8	3.3
Urban	384	96.7	100.0

Figure 4.3 Histogram of branch area type (br_area_type) descriptive variable.

The data exploration phase highlighted two variables, the *number of mass market customers* (br_primary_markets) and *branch area type* (br_area_type), that can be removed from further analysis. The following section will make additional transformations to variables to be used during CA. A final list of variables selected for use can be seen in Table 4.5.

4.3.3 Modifying and transformation of data

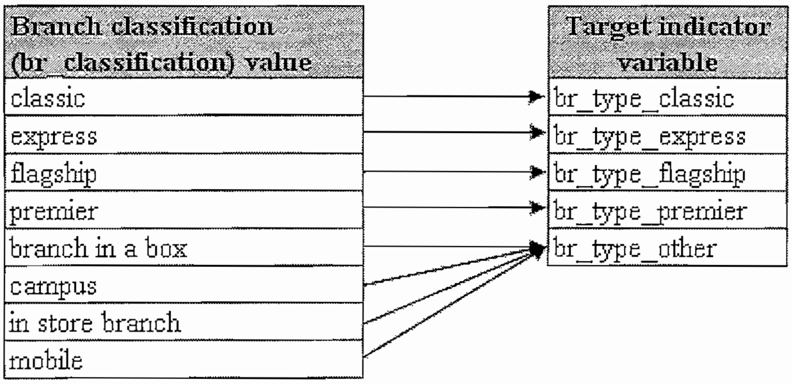
During the modification and transformation of variables, they are cleaned and transformed in a manner that will improve the data mining results. The following three sections will cover the creation of indicator variables, standardisation of variable values and inclusion of domain knowledge by means of weighted variables.

4.3.3.1 Creating indicator variables for the character variables

Branch classification is a character value. This poses a problem as the *k*-means algorithm needs numerical values. A possible solution to this problem would be to create a numeric variable and assign a number to each distinct value of the variable. For example, a classic branch would then receive any number from one to eight for an "in store" branch classification. This, according to Berry and Linoff (2000:554), is a naïve approach. The problem is that if there are 8 distinct values, as is the case with this variable, the branch category to which the last value is assigned, will be deemed more important than the first value.

A superior approach to solve this problem would be to create a set of indicator variables, one for each distinct value of the variable. Thus, if in the case of the branch classification variable (br_classification) a branch is classified as a classic branch, the classic branch indicator variable (br_type_classic) will be set to one, while the other remaining branch indicator variables will be set to zero. Albeit the fact that this approach increases the number of variables, it removes the problem of biasness associated with the previous approach by neutralising the implication of ordering among the labels.

An additional advantage of creating a set of indicator variables is that it allows the modeller liberty to include domain knowledge during the mapping phase (Pyle, 1999:195). Branches classified as branch in a box, campus, in store and mobile branches are predominantly in a developmental phase with a relatively few number of branches operating. It was decided to map these branches into a single indicator variable. Table 4.4 shows the mapping of the classification to target indicator variables.



Branch classification (br_classification) value	Target indicator variable
classic	br_type_classic
express	br_type_express
flagship	br_type_flagship
premier	br_type_premier
branch in a box	br_type_other
campus	br_type_other
in store branch	br_type_other
mobile	br_type_other

Table 4.4 Mapping of branch classification (br_classification): A character value to a numerical equivalent indicator variable.

4.3.3.2 Standardising data in order to create comparable ranges

Variables analysed are measured in different units and over different ranges. For example, a branch has hundreds of customers and thousands of accounts. The implication of this is that the number of accounts will be more significant than the number of customers as a result of the difference in magnitude of numbers.

Standardisation may be very useful when using techniques that perform mathematical operations directly on the values, such as in the case of *k*-means clustering (Berry & Linoff, 2004:23). It was decided to standardise the variables to a mean of 0 and a standard deviation of 1. This effectively means that all variables will contribute equally when the distance between two records is computed. All variables used as part of the CA were thus standardised.

4.3.3.3 Apply domain knowledge by means of weighting more important variables

In the previous section, variables were standardised to allow all variables to equally contribute during the clustering phase. In some instances, the inverse is also required as some variables must perform more dominantly during analysis. Applying domain knowledge is accomplished by assigning a weight to the desired variables. Management felt that the current branch classification (br_classification) was a good indicator of branch activity, therefore the branch classification variables (br_type_classic, br_type_express, br_type_flagship, br_type_premier and br_type_other) were adjusted to a weight of ten.

4.3.4 Clustering analysis modelling

The preceding section covered the data preparation phase required prior to applying CA. This section will cover the modelling phase applied during CA.

The objective of applying CA prior to DEA is to assure that branches compared share a high level of similarity. This ultimately will guarantee that when a branch is identified as efficient, by DEA, it will not be due to any biasness in the data, but rather as a result of the branch truly out performing another on merit. In order to assure a satisfactory level of homogeneity among branches to be compared, it was decided to apply two iterations of CA.

The first iteration of CA focused on removing outliers. This was accomplished through applying CA to the branch data and selecting clusters that are in a close proximity of one another (having a relative level within the cluster similarity), but excluding clusters with a relatively low count of branches. As a result of focusing more on the removal of outliers, it might be possible that clusters of inferior quality could be selected.

The second iteration was applied to the clusters identified as being appropriate during the first iteration of CA. This iteration will focus exclusively on the creation of high quality clusters.

4.3.4.1 Software used for clustering analysis

SAS 5.2 Enterprise Miner software package was used to design and test different models for solving the clustering problem. SAS perform CA by means of a procedure called PROC FASTCLUST. This procedure finds clusters of objects, in this case homogeneous branches of the financial institution, by using the *k*-means clustering algorithm (Fernandez, 2003:93). Creating a data model consists of defining a data source, in this case the branch data set and specifying clustering rules applicable to the clustering algorithm.

4.3.4.2 Variables used during clustering analysis

The variable describing the area a branch operates in (br_area_type) and the number of mass market customers (br_primary_Markets) was eliminated from the analysis on account of reasons provided in Section 4.3.2. Descriptive variables, for instance branch name, etc. presented in Table 4.1, were removed from the analysis data set as previously mentioned in Section 4.3.

New variables in the data set include the branch classification variables (br_type_classic, br_type_express, br_type_flagship, br_type_premier and br_type_other) created during the data modification and transformation stage (Section 4.3.3). Table 4.5 lists the final set of variables used during the CA. Variables 7 to 16 have been standardised to a mean value of 0 and a standard deviation of 1.

Variable number	Variable	Type	Description
1	Br_Num	Numeric	Uniquely identifies branch
2	br_type_classic	Numeric	Indicate a classic branch category
3	br_type_express	Numeric	Indicate an express branch category
4	br_type_flagship	Numeric	Indicate a flagship branch category
5	br_type_premier	Numeric	Indicate a premier branch category
6	br_type_other	Numeric	Indicate branch category as being one of the following: branch in a box, campus, in store branch or mobile branch
7	stnd_br_FreqABF	Numeric	Number of assets based accounts per branch
8	stnd_br_FreqCA	Numeric	Number of current accounts per branch
9	stnd_br_FreqCAMS	Numeric	Number of credit card accounts per branch
10	stnd_br_FreqHL	Numeric	Number of home loan accounts per branch
11	stnd_br_FreqINV	Numeric	Number of investment accounts per branch
12	stnd_br_FreqMM	Numeric	Number of money market accounts per branch
13	stnd_br_FreqSA	Numeric	Number of savings accounts per branch
14	stnd_br_Primary_CG1	Numeric	Number of affluent customers per branch
15	stnd_br_Primary_CG2	Numeric	Number of high valued customers per branch
16	stnd_br_Primary_SmallBus	Numeric	Number of small business customers per branch

Table 4.5 Variables used as input for *k*-means clustering.

4.3.4.3 Clustering as a tool for removing outliers

Good clustering results depend on a k-value that corresponds to the natural structure of the data. This value specifies the number of clusters a data set is to be clustered into and is specified beforehand. Determining the correct number of clusters may be difficult and is considered a major drawback when using the *k*-means clustering algorithm (Berkhin, 2002:17). Because it is difficult to determine an appropriate k-value, it was decided to run the clustering process several times with different k-values.

The first run made use of automatic cluster detection using cubic clustering criterion (CCC) to automatically determine the best number of clusters. CCC is a method used to estimate the number of clusters when using *k*-means or other methods based on minimising the within-cluster sum of squares. CCC values greater than 2 and 3 indicate good clusters; values between 0 and 2 indicate potential clusters, but Fernandez (2003:134) warns that these clusters should be evaluated carefully.

In successive runs, the k-value was increased by one starting at an initial value of three. Results were analysed to determine the k-value producing the best clustering results. Table 4.6 displays the results obtained by changing the k-value in the clustering process configuration of six consecutive runs. Variables in Table 4.6 describing the clustering results are the following:

- k-value – a number specifying the amount of clusters to be created;
- cluster number;
- cubic clustering criterion (CCC) – measure of cluster quality;
- number of branches per cluster;
- root mean square across variables of the cluster standard deviations (RMSSTD), an indication of the within cluster variability. A small value is a good sign of a homogeneous cluster and will be the primary measure of cluster quality in this study;
- maximum distance from cluster seed. A smaller value indicates that the branches are in a close proximity of the cluster seed, indicating closely knit clusters and will be the secondary measure of cluster quality in this study;
- nearest cluster;
- distance to the nearest cluster.

From Table 4.6 it can be seen that auto cluster detection, using the CCC for measuring cluster quality, created two clusters with a CCC value of 1.63005. Although the CCC value was higher than any of the consecutive runs, only two clusters were created with a significantly large RMSSTD (root mean square across variables of the cluster standard deviations) value of 1.70406 for auto_1 and 1.10624 for auto_2. These clusters, when investigated, also contained branches from different branch classifications.

Further analysis revealed that across different k-values of 3, 4, 5 and 6, one large cluster containing 272 branches, all of which were classified (br_type_classic) as classic branches, was present. This one large cluster stayed stable for different k-values. Statistics, i.e. number of branches in cluster, root mean square across variables of the cluster standard deviations (RMSSTD) and maximum distance from cluster seed, describing the clusters 3.01, 4.04, 5.05 and

6.04, also stayed consistent. The only change noted was that some of the smaller clusters were separated. For example, from a k-value of 3 to a k-value of 4, cluster 3.03 was divided into two smaller clusters, namely 4.01 and 4.03. Analysis showed that clusters 4.01 and 4.03 combined were exactly the same as cluster 3.03.

K-value	Cluster number	Cubic clustering criterion (CCC)	Number of branches in cluster	Root mean square across variables of the cluster standard deviations (RMSSTD)	Maximum distance form cluster seed	Nearest cluster	Distance to nearest cluster
Auto Cluster	Auto_1	1.6301	342	1.7041	15.1545	2	12.9723
	Auto_2		56	1.1062	12.9934	1	12.9723
3	3.01	1.1813	272	0.7172	11.6480	3	11.6317
	3.02		56	1.1062	12.9934	3	11.7911
	3.03		70	2.2479	12.8583	1	11.6317
4	4.01	0.9881	45	1.9683	12.3657	4	12.4320
	4.02		56	1.1062	12.9934	1	12.5274
	4.03		25	0.6901	6.5808	1	12.4754
	4.04		272	0.7172	11.6480	1	12.4320
5	5.01	0.7871	16	0.8829	7.3263	2	14.1587
	5.02		56	1.1062	12.9934	1	14.1587
	5.03		29	0.7728	6.4834	5	14.1919
	5.04		25	0.6901	6.5808	5	14.1664
	5.05		272	0.7172	11.6480	4	14.1664
6	6.01	0.7397	16	0.8829	7.3263	5	14.1868
	6.02		11	1.3261	9.4550	5	6.9891
	6.03		29	0.7728	6.4834	4	14.1919
	6.04		272	0.7172	11.6480	5	14.1544
	6.05		45	0.6898	5.3774	2	6.9891
	6.06		25	0.6901	6.5808	4	14.1664
7	7.01	0.6368	16	0.8829	7.3263	7	14.1868
	7.02		11	1.3261	9.4550	7	6.9891
	7.03		29	0.7728	6.4834	6	14.1969
	7.04		65	0.8717	9.7174	5	4.1352
	7.05		207	0.4085	4.1260	4	4.1352
	7.06		25	0.6901	6.5808	5	14.1514
	7.07		45	0.6898	5.3774	2	6.9891

Table 4.6 Results of clustering iteration1, the removal of outliers.

The behaviour of one large cluster being sub-divided into smaller clusters remained constant until a k-value of 7 was specified. The only difference between separating the data set into 6 and 7 clusters respectively, was that cluster 6.04 was separated into two smaller clusters, namely 7.04 and 7.05. Analysis again showed that clusters 7.04 and 7.05 combined were equivalent to cluster 6.04.

From the above, it is evident that different k-values did not necessarily improve the clustering results. In view of these findings it was decided to use the clusters created with a k-value of three. Since, as the k-value increased, the only difference noted, was the fact that it was mostly the smaller clusters that were separated. Cluster 3.01, created with a k-value of 3 and consisting of classic branches only, was the only one that stayed consistent with different k-values. Branches in cluster 3.01 will be used for the remainder of this research. Figure 4.4 graphically displays statistics when the data set was clustered into three clusters.

From Figure 4.4, it is easy to note that the branch classification variables (br_type_classic, br_type_express, br_type_flagship, br_type_premier and br_type_other) played a dominant part during the cluster creation process. This behaviour is due to the fact that the branch classification variables were assigned extra weight. The result was in line with what was expected.

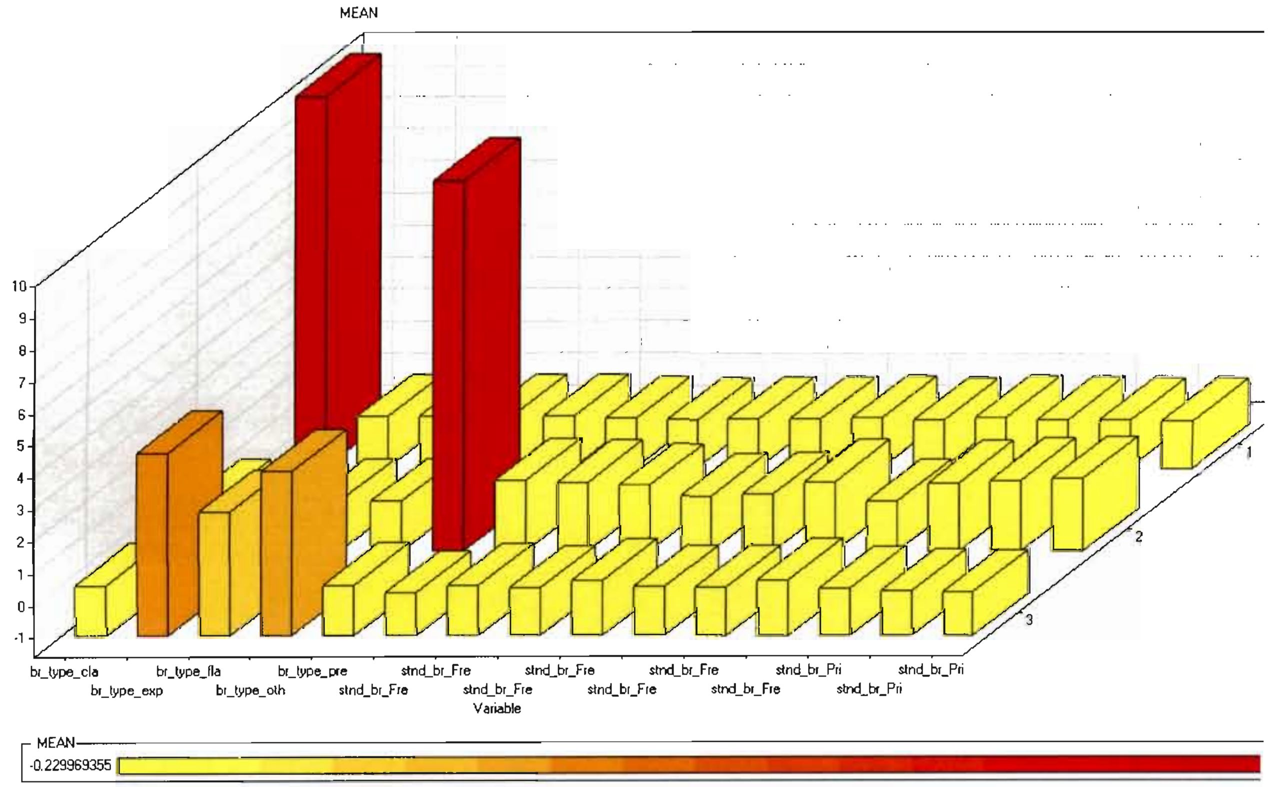


Figure 4.4 Iteration 1 clustering results. Important variables for discriminating between clusters can be seen.

The following section covers the second iteration of CA. During this iteration, the focus will move from outlier removal to the creation of high quality clusters meeting the criteria set by the research during the clustering phase.

4.3.4.4 Selecting homogeneous bank branches

The criteria for clustering were the selection of a group of branches sharing a high level of similarity. In order to meet the set criteria, it was decided to apply two iterations of CA. The first iteration, dealt with in the previous section, focused on the removal of outliers, while the second iteration is dedicated to the creation of high quality clusters as will be covered in this section.

A high level of similarity shared by branches selected was not the only criteria to be met. In addition, a sufficient number of similar branches had to be selected allowing for meaningful analysis in the DEA and DTI phases. Thus, the CA done in the second iteration will be a trade-off between high cluster quality and the selection of a satisfactory number of branches.

The first iteration of CA produced a group of branches, all of which were classified as classic branches (br_type_classic). Therefore, since all branches post the first iteration were classic branches, variables describing the branch classification were removed from further analysis. Table 4.7 lists the variables used during the second iteration of CA. All variables, except the branch number (Br_Num) variable listed in Table 4.7, were standardised to a mean value of 0 and a standard deviation of 1. The actual data used as input during the second phase of CA is listed in Appendix A.

N	Variable	Type
1	Br_Num	Numeric
2	stnd_br_FreqABF	Numeric
3	stnd_br_FreqCA	Numeric
4	stnd_br_FreqCAMS	Numeric
5	stnd_br_FreqHL	Numeric
6	stnd_br_FreqINV	Numeric
7	stnd_br_FreqMM	Numeric
8	stnd_br_FreqSA	Numeric
9	stnd_br_Primary_CG1	Numeric
10	stnd_br_Primary_CG2	Numeric
11	stnd_br_Primary_SmallBus	Numeric

Table 4.7 Variables used in the second iteration of CA.

Akin to the first iteration, it was decided to run CA with different k-values for the 272 branches of cluster 3.01 formerly created. While using the cubic clustering criterion (CCC) as method to determine the appropriate number of clusters, the k-values were manually incremented by one. Different k-value clustering results were then analysed to determine which k-values produced the best results.

K-value	Cluster number	Cubic clustering criterion (CCC)	Number of branches in cluster	Root mean square across variables of the cluster standard deviations	Maximum distance form cluster seed	Nearest cluster	Distance to nearest cluster
6	6.01	0.4770	51	0.5900	3.0918	6	2.2070
	6.02		145	0.3449	2.6141	6	2.5325
	6.03		12	0.6767	2.7537	6	3.9592
	6.04		1	0.0000	0.0000	3	7.2602
	6.05		11	0.8560	3.8508	1	4.5109
	6.06		52	0.5422	3.2870	1	2.2070
7	7.01	0.4581	17	0.7779	3.6461	6	3.5675
	7.02		113	0.2898	2.6504	5	1.8890
	7.03		58	0.5289	3.1589	5	1.9470
	7.04		21	0.7371	3.3717	3	3.9305
	7.05		59	0.4243	3.1992	2	1.8890
	7.06		3	0.6122	2.0787	1	3.5675
	7.07		1	0.0000	0.0000	4	8.0335
8	8.01	0.4334	3	0.6122	2.0787	3	3.3834
	8.02		36	0.5572	3.0030	7	2.2266
	8.03		11	0.7753	3.1225	1	3.3834
	8.04		28	0.4927	2.2848	2	2.2804
	8.05		108	0.2733	2.1432	7	1.8181
	8.06		14	0.6914	3.0270	2	3.5419
	8.07		71	0.4261	3.3631	5	1.8181
	8.08		1	0.0000	0.0000	6	7.4314
9	9.01	0.4255	8	0.7087	2.5565	3	3.2358
	9.02		108	0.2733	2.1432	9	1.8181
	9.03		3	0.5515	1.7577	1	3.2358
	9.04		30	0.5206	2.5935	6	2.2420
	9.05		3	0.6122	2.0787	3	3.5874
	9.06		34	0.5391	2.9731	9	2.1511
	9.07		14	0.6914	3.0270	6	3.5533
	9.08		1	0.0000	0.0000	7	7.4314
	9.09		71	0.4261	3.3631	2	1.8181
10	10.01	0.4041	6	0.7041	2.5577	5	3.1951
	10.02		37	0.4541	3.2298	8	1.8432
	10.03		14	0.5810	2.7170	2	2.7724
	10.04		115	0.2855	2.1821	8	1.9459
	10.05		3	0.6122	2.0787	1	3.1951
	10.06		4	0.5395	1.7456	9	3.1573
	10.07		31	0.4646	2.3812	2	1.8949
	10.08		46	0.4100	3.0228	2	1.8432
	10.09		15	0.6078	2.4948	2	3.0592
	10.1		1	0.0000	0.0000	6	7.0327
11	11.01	0.3986	113	0.2791	2.1775	7	1.9129
	11.02		40	0.4533	3.1983	7	1.7733
	11.03		31	0.4646	2.3812	2	1.9173
	11.04		1	0.0000	0.0000	5	7.0327
	11.05		4	0.5395	1.7456	9	3.1573
	11.06		2	0.6786	1.5174	8	3.3942
	11.07		44	0.4064	3.0319	2	1.7733

K-value	Cluster number	Cubic clustering criterion (CCC)	Number of branches in cluster	Root mean square across variables of the cluster standard deviations	Maximum distance form cluster seed	Nearest cluster	Distance to nearest cluster
	11.08		5	0.6173	2.0536	10	2.9710
	11.09		15	0.6078	2.4948	2	3.0882
	11.1		3	0.5515	1.7577	8	2.9710
	11.11		14	0.5723	2.6885	2	2.7584
12	12.01	0.3835	3	0.4831	1.6181	4	3.5935
	12.02		1	0.0000	0.0000	1	7.0926
	12.03		15	0.4191	1.7098	12	2.0659
	12.04		12	0.6038	2.7222	6	2.6075
	12.05		48	0.3949	2.2860	12	1.7730
	12.06		14	0.5490	2.2842	12	2.2753
	12.07		5	0.6173	2.0536	9	2.9710
	12.08		109	0.2650	1.5502	5	1.8944
	12.09		3	0.5515	1.7577	7	2.9710
	12.1		2	0.6786	1.5174	7	3.3942
	12.11		13	0.5686	2.6947	12	2.6758
	12.12		47	0.4107	2.0588	5	1.7730
13	13.01	0.3728	14	0.4986	2.2619	7	2.0351
	13.02		33	0.4013	2.7608	6	1.6579
	13.03		104	0.2519	1.5878	6	2.0959
	13.04		4	0.5395	1.7456	5	3.1596
	13.05		10	0.5489	1.9046	13	2.5673
	13.06		61	0.3965	1.7937	2	1.6579
	13.07		14	0.5771	3.0519	1	2.0351
	13.08		1	0.0000	0.0000	4	7.0327
	13.09		15	0.4161	1.7821	6	2.1876
	13.1		3	0.6122	2.0787	12	3.4934
	13.11		3	0.5515	1.7577	12	3.0514
	13.12		3	0.4866	1.4401	11	3.0514
	13.13		7	0.4641	1.9144	7	2.2996

Table 4.8 Results of clustering iteration 2: The selection of homogeneous branches.

Results obtained from consecutive clustering runs with only the k-value being incremented are displayed in Table 4.8. During these attempts it was noticed that one large cluster consistently materialised. The k-value affects the number of clusters a data set was clustered into. This homogeneous group of branches consistently ended in the same cluster and in addition, constantly had a lower root mean square across variables of the cluster standard deviations (RMSSTD) value and across different k-values. From this observation, it became clear that the data set contained a large number of branches sharing a high level of similarity.

Auto cluster detection, using the CCC method to cluster the branch data set into 12 clusters, was selected for further analysis. The selection of auto clustering results was based on the following findings:

- One large cluster (12.08) was present consisting of 109 branches, sufficient to perform a DEA;
- cluster 12.08 had a root mean square across variables of the cluster standard deviations (RMSSTD) value of 0.26501, which was the second lowest RMSSTD value compared to other k-value results, indicating a relatively low within cluster standard deviation;
- branches in cluster 12.08 had the smallest radius from the cluster seed signifying that all the branches of this cluster were in close proximity of the cluster seed.

4.3.4.5 Analysis of branch data clustered into twelve clusters

The results of the final clustering iteration were presented to management to determine whether the clustering results were in accordance with management expectations. The inspection of results would purely be an objective interpretation based on first hand experiences with the branches.

To simplify the evaluation of clustering results for management, it was decided to use the branches clustered together in cluster 12.03. Cluster 12.03 was selected as it contains only fifteen branches, the majority of which are well known branches located within the province.

Management were pleased with the branches and reached consensus that ten of the fifteen branches do, as a matter of fact, share a high level of similarity (greyed branches in Table 4.9). Furthermore, they were enlightened to find the Hermanus (345) and Hout bay branch (676) in this group.

Branch number	Branch name	Town	Province	Branch distance from cluster seed
48	Rondebosch	Cape Town	Western Cape	1.3426
224	Nelson Mandela Square	Sandton	Gauteng	1.4952
310	Ballito	Ballito	Kwazulu Natal	1.3371
328	Broadacres	Randburg	Gauteng	1.4355
343	Cascades	Pietermaritzburg	Kwazulu Natal	0.9723
345	Hermanus	Hermanus	Western Cape	1.2730
353	Umhlanga Rocks	Durban	Kwazulu Natal	0.6887
515	Musgrave	Durban	Kwazulu Natal	1.3162
627	Preller Square	Bloemfontein	Free State	1.6178
676	Hout Bay	Cape Town	Western Cape	0.7873
958	Rosebank	Johannesburg	Gauteng	1.3266
971	Bryanston	Sandton	Gauteng	1.2109
973	Benmore	Sandton	Gauteng	1.7098
976	Illovo	Sandton	Gauteng	1.2665
977	Rosebank Mall	Johannesburg	Gauteng	0.9696

Table 4.9 Branches from cluster 12.03 sharing a high level of similarity.

Based on management's analysis and approval of cluster 12.03, it was concluded that the clustering employed, produced excellent results and that cluster 12.08, selected in Section 4.3.4.5, can therefore be used in the following sections as a group of similar branches.

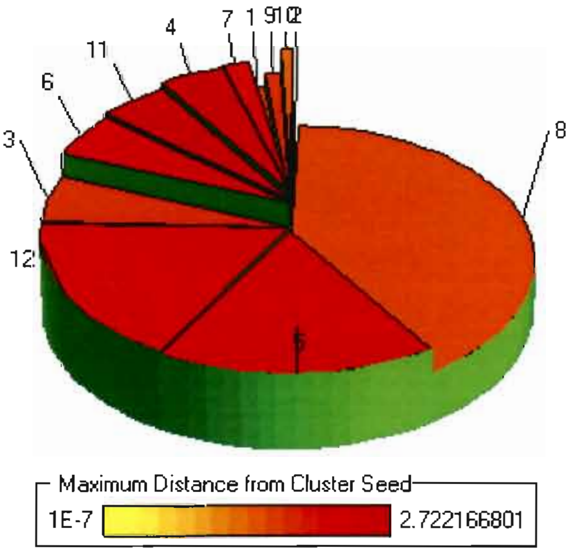


Figure 4.5 Pie chart displaying12 branch clusters.

Figure 4.5 displays a pie chart summarising 3 important statistics relating to the clustering results. The width of the pie slice represents the number of branches in the cluster, standard deviation is represented by the pie slice height, while colour represents the cluster radius with red, a large radius, indicating that branches are spread distantly around the cluster seed. For example, cluster eight, the biggest slice of the pie, had the most branches, as well as the lowest standard deviation. Clusters five and twelve, according to the graph, are reasonably similar being similar in pie slice, height and colour. Table 4.10 displays the importance of the variables in CA.

Variable	Importance
Number of current accounts (br_freqCA)	1
Number of home loan accounts (br_freqHL)	0.9497
Number of small business customers (br_primary_SmallBus)	0.5521
Number of affluent customers (br_primary_CG1)	0.5306
Number of savings accounts (br_freqSA)	0.5076
Number of credit card accounts (br_freqCAMS)	0.5055
Number of high value customers (br_primary_CG2)	0.3663
Number of asset based finance accounts (br_freqABF)	0
Number of investment accounts (br_freqINV)	0
Number of money market accounts (br_freqMM)	0

Table 4.10 Importance assigned to variables during clustering.

Table 4.10 lists all the variables used as input for the clustering algorithm. An importance value between 0 and 1 is calculated for every variable in the input data set. If a variable displays a value of 1, it means that the variable played an important part during the cluster construction. From Table 4.10, it can be seen that current accounts (br_freqCA) and home loan accounts (br_freqHL) were the variables mostly influencing the CA, while the number of asset-based finance accounts (br_freqABF), investment accounts (br_freqINV) and money market accounts (br_freqMM) did not affect the clustering results. This does not necessarily mean that these variables can be removed from the CA. Figure 4.6 graphically displays the layout of the clusters according to the variables, only the variables playing a major part in the CA are shown in this figure.

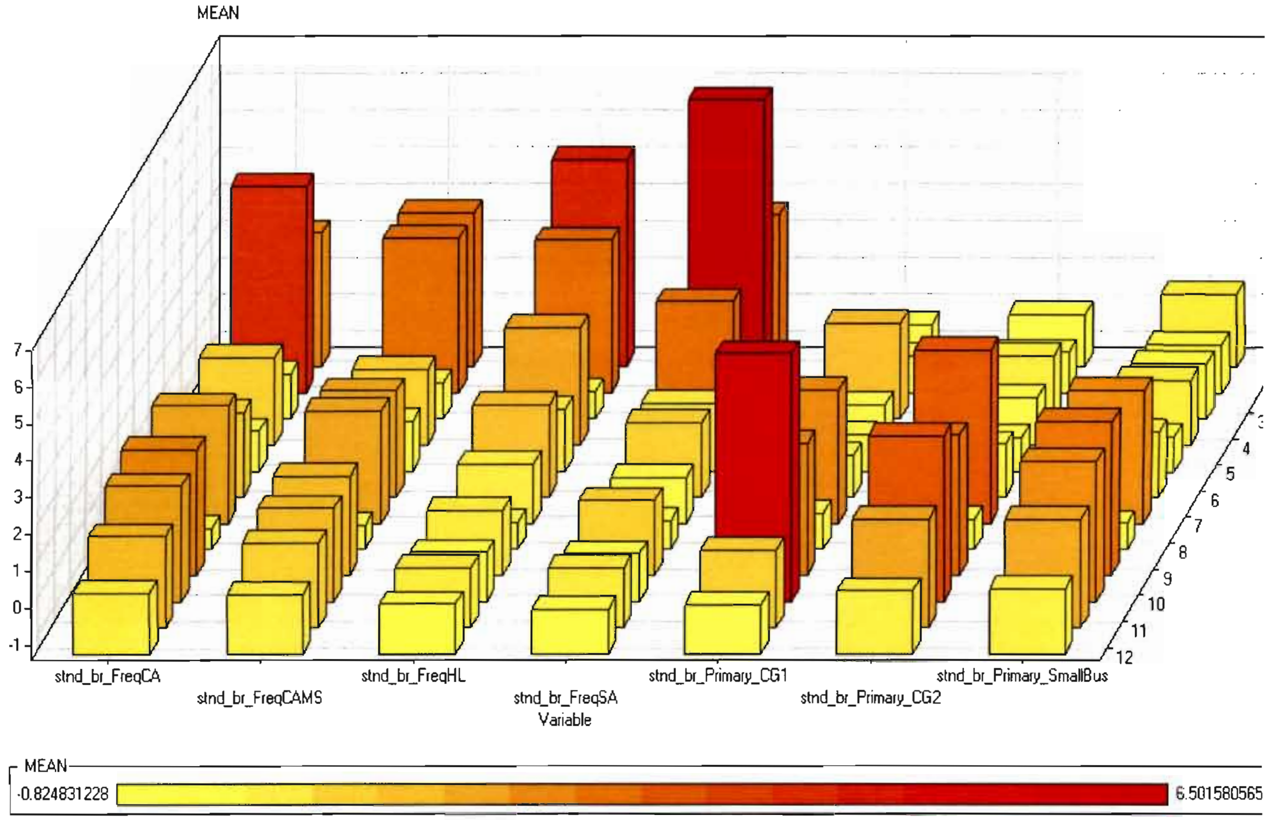


Figure 4.6 Graphic display of the differences between clusters according to variables.

From Figure 4.6, it is observable that cluster 12.08, consisting of the most branches, on average has less accounts and fewer affluent and high value customers (br_primary_CG1, br_primary_CG2) than all the other clusters. On further analysis, it came to light that cluster eight consisted primarily of average market customers (br_Primary_Market). Thus, branches in cluster 12.08 are mainly focused on the average customer. On account of analyses done, the 109 branches identified as being similar in cluster 12.08, were selected for further use in the research.

4.4 Chapter conclusion

The focus of this chapter was to identify branches that are homogeneous with regards to certain criteria. As branches within the financial institution are currently segmented, the chapter started with three convincing arguments for improving the segmentation. The chapter proceeded by briefly discussing the SEMMA methodology as proposed by the SAS Institute.

The SEMMA methodology was applied and the data provided, clustered accordingly. CA started with 398 retail branches, and a first iteration was applied to the entire branch data set. The purpose of this iteration was to remove outlier branches resulting in 272 branches selected. A second iteration of CA was conducted, this time focusing on the creation of high quality clusters. Finally, 109 homogeneous branches, based on eleven variables, were selected.

The following chapter will use the 109 branches identified as being similar and apply DEA to these branches.

Chapter 5 - Measuring relative branch efficiency

5.1 Introduction

Data envelopment analysis (DEA) is the second in a three step analytical process and is a method to comprehensively measure the relative efficiency of a decision making unit (DMU). In this study a branch of the financial institution represents a DMU. Figure 5.1 depicts the second step of the analytical model consisting of three steps. The discriminatory power of DEA could severely be compromised if a high level of heterogeneity exists within the data under investigation (Samoilenko & Osei-Bryson, 2008:1568). Consequently, it was decided to use clustering analysis (CA), described in Chapter 4, as a pre-process to DEA to ensure homogeneity among the branches analysed. This resulted in 109 branches identified as being homogeneous and will be used for DEA.

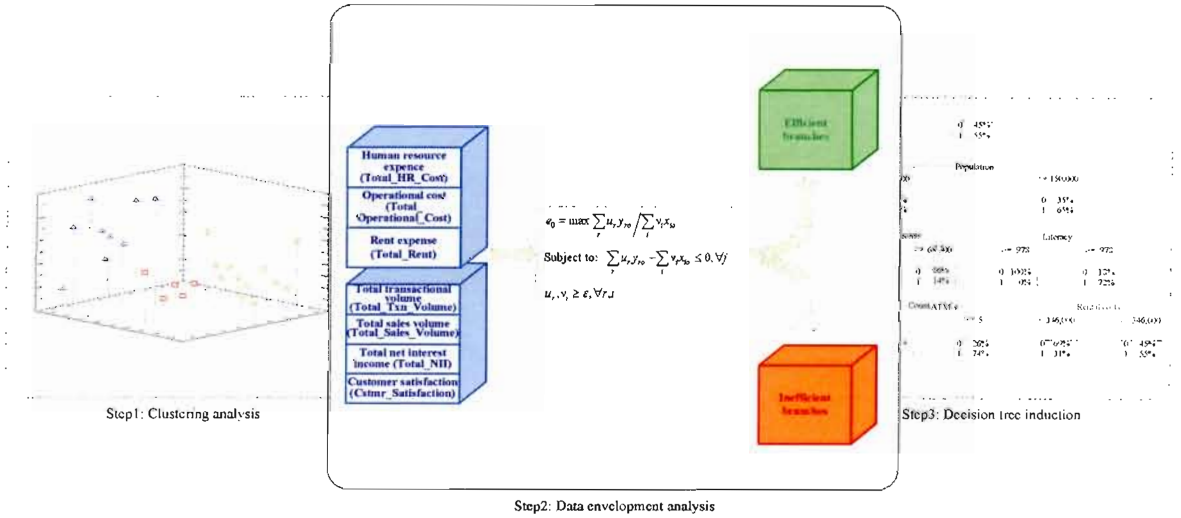

Step1: Clustering analysis
Human resource expence (Total_HR_Cost)
Operational cost (Total Operational_Cost)
Rent expense (Total_Rent)
Total transactional volume (Total_Txn_Volume)
Total sales volume (Total_Sales_Volume)
Total net interest income (Total_NII)
Customer satisfaction (Cstmr_Satisfaction)
$e_0 = \max \sum_r u_r y_{ro} / \sum_i v_i x_{io}$
Subject to: $\sum_r u_r y_{ro} - \sum_i v_i x_{io} \leq 0, \forall j$
$u_r, v_i \geq \varepsilon, \forall r, i$
Efficient branches
Inefficient branches
Step3: Decision tree induction
Step2: Data envelopment analysis

Figure 5.1 The second of a three step analytical model, DEA.

In terms of this research, the results generated from DEA will be used to serve two objectives. Firstly, the results will indicate branches operating efficiently, and these will be used in the third step of the analytical process, decision tree induction (DTI). Secondly, the results will be used as a benchmark for improving operating efficiencies of inefficient branches.

5.2 Data envelopment analysis procedures and characteristics

The past twenty years saw an increase in popularity of DEA as a tool for determining the relative efficiency of a set of decision making units (Mostafa, 2008:311). An overview of DEA was given in Chapter 3, but certain characteristics of DEA will be mentioned once more to link them to the empirical study.

The main elements in DEA are a set of decision making units (DMUs), which in this study are branches of a financial institution, together with their inputs and outputs. DMUs may portray any entity as long as they utilise the same inputs to produce the same outputs. DEA measures efficiency by estimating an empirical production function expressing the highest values of benefits (outputs) that could be generated from given resources (inputs).

The Charnes, Cooper and Rhodes (CCR) and Banker, Charnes and Cooper (BCC) DEA models are the two basic DEA models. The CCR model assumes constant returns to scale (CRS), whereas the BCC model presumes variable returns to scale (VRS). This study applies DEA based on the CCR model with the understanding that branches being measured are homogeneous, therefore economies of scale minimal, if at all present. The CCR and BCC models may be expressed as either input-oriented, or output-oriented depending on the efficiency being measured. The input-oriented model minimises inputs with given outputs, whereas the output-oriented model maximises outputs with given inputs. In this study, the output-oriented version of the CCR model was applied, and therefore the maximum outputs a branch can produce, given certain inputs, will be measured. Microsoft Excel software package was used to compute the efficiency analysis.

5.3 Variable selection and identification for data envelopment analysis

The DEA model is based on the inputs and outputs describing a decision making unit. Consequently, when conducting DEA, the first step is to identify essential input and output variables, while selected variables must associate with the efficiency being assessed. For example, when measuring efficiency of decision making units producing milk, it will not make sense to include variables describing the surrounding area, such as average population, etc. It will however make sense to include variables like number of cows and litres of milk produced. By vigilantly selecting variables, nearly any type of efficiency can be measured, provided selected variables relate to the efficiency being measured. Measuring efficiency with DEA has the advantage of selecting measures not necessarily commensurable, meaning that number of customers may be compared against number of transactions, etc., thus allowing freedom to select nearly any comparable measure.

The efficiency assessed relates directly to the input and output measures selected and therefore the selection of a large number of variables are very alluring. Baker (2006:174) warns that a large number of variables have the propensity of causing nearly every unit to appear efficient and recommended that the number of DMUs should at least be equal to two or three times the number of inputs and outputs. In this study, a total of seven variables will be used to measure 109 branches, which is significantly more than the minimum recommended.

In the literature there is no specific agreement on how to measure efficiency of a financial institution. However, two general approaches, the intermediation and production approaches exist (O'Donnell & Van Der Westhuizen, 2002:500). The intermediating approach views financial institutions as a mediator between investors and savers. This approach assesses assets and equity and tries for example, to maximise measures akin to rate on assets (ROA) and rate on equity (ROE).

The DEA model will follow an approach corresponding to that of the production approach. According to Mostafa (2008:314), the production approach views a financial institution as a firm delivering services to its customers in the form of transactions. Efficiency, in accordance with the production approach, measures the extent to which a financial institution combines and utilises resources to produce, among others, the maximum amount of transactions. In this study, the intention is aimed at maximising outputs relevant to transactions with inputs, such as operating expenses given. Thus, it was decided, as a starting point, to review various resources a typical branch would utilise. Management familiar with branch maintenance was approached to assist with the identification of important branch characteristics relating to operating efficiency and in line with the production approach.

5.3.1 Input variable selection, motivation and description

Management was cooperative in assisting, and thirteen potentially influential resources (inputs) were identified. Although this list was not exhaustive by any means, and many more variables could have been included, they believed these variables to be adequate. It was, however, decided to reduce the total number of variables relating to resources, considering the fact that branch resources are practically divided into three main categories, namely human resources, operational and rent related expenses. For example, computer equipment and building expenses logically form part of maintenance expenses, which in turn is a component of daily operational cost incurred by a branch. The thirteen sources identified were carefully mapped to one of these overarching categories. Table 5.1 lists the hierarchy used to derive the final input and output variables.

Human resource is acknowledged as one of the major resources utilised by branches (Fatti & Clarke, 1998:57; Vassiloglou & Giokas, 1990:592). Variables such as administrative salaries, bonuses, overtime, contract workers and part time tellers were mapped to human resource related expenses. Thus, total human resource expense (Total_HR_Cost) was the sum of these five variables per branch.

Resources (Inputs)	Human resource expense (Total_HR_Cost)	Administrative salaries	
		Bonuses paid out	
		Overtime paid	
		Contract workers paid	
		Part time tellers paid	
	Operational cost (Total_Operational_Cost)	Maintenance	Computer equipment
			Building
		Stationery used	
		Direct marketing	
		Post office charges	
		Security guards	
		Telephone account	
	Rent expense (Total_Rent)	Rent expense / Transfer cost pricing	
Benefits (Outputs)	Total transactional volume (Total_Txn_Volume)		
	Total sales volume (Total_Sales_Volume)		
	Net interest income (Total_NII)		
	Customer satisfaction (Cstmr_Satisfaction)		

Table 5.1 Input and output variables.

Operational cost is the sum of six variables selected by management as representative of general operational cost, i.e. day to day expenses a branch would incur.

Rent expense is a good thermometer of the commercial worth of a location, as affluent market environments relate to higher rental rates. Despite the fact that rent is a fixed cost over which management has little or no leverage, the decision was made to include rental expenses in the analysis. As branches located in a building owned by the financial institution are liable to reimburse the financial institution the fair value of the property, this value was set to be the minimum market value of the property. In such cases, transfer price costing substitutes the rental value, elsewhere the rent expenditure of the branch to an external party is used as the rental cost.

5.3.2 Output variable selection, motivation and description

In utilising inputs, financial institutions presuppose branches to generate benefits (outputs). The production approach, as applied in this study, measures how well branches convert selected inputs, such as human resource cost, operational cost and rental expenses into outputs. Total transactional volume (Total_Txn_Volume) was the first output selected by management and represents total number of transactions conducted by tellers. Similar studies tend to divide transactions into deposits, loans and investment transactions. Management's view at this financial institution, however, was that retail branch tellers offer identical services to customers and therefore will holistically measure all types of transactions as a single measure (Mostafa, 2008:310).

Sales volume per branch (Total_Sales_Volume) was identified as the second output measure and embodies total volume of products sold. Once again, management felt that retail branches offer the same products to customers and hence decided to represent sales as a single measure.

Net interest income (Total_NII) and customer satisfaction (Cstmr_Satisfaction) fall outside the scope of the production approach. Net interest income is the difference between interest paid on deposits and interest earned with loans, and this measure is typically used in the intermediation approach. Net interest income is a good indicator of a branch's capability to mediate between surplus funds (savers) and short fall of funds (investors) and was therefore included.

The financial institution is well known as a customer centred concern. High output per branch is required, but not at the expense of the customer, therefore customer satisfaction formed part of the efficiency analysis. A survey was conducted rating the financial institution from a customer's perspective. The questionnaire comprised several sections covering different aspects of the financial institution, such as product related questions, etc. The score obtained from the customer satisfaction section was used in this study. Table 5.2 lists customer satisfaction related questions used to calculate the level of satisfaction. The rating a branch obtained is a value between 0% and 100% as rated by customers transacting at a branch.

Question number	Question
1	Being treated with care
2	Courteous and friendly staff
3	Clear and easily understood communication
4	Staff conveying trust and confidence
5	Personal service that matches your specific needs
6	Easily approachable and accessible branch staff
7	Pleasant appearance of the branch and staff
8	Fast and efficient processes
9	Waiting time in queues being managed well
10	Consistently delivering on promises made
11	Staff quick to respond and willing to help
12	Feeling safe and secure in the branch
13	Staff conduct – showing respect, integrity and excellent service
14	Knowledgeable and competent staff
15	The manner in which the security officer communicates and the image he/she portrays

Table 5.2 Questionnaire used to calculate level of satisfaction.

This section covered variables selected for use during efficiency analysis. The selection of appropriate variables is extremely important, as it will establish the nature of efficiency being assessed. The following section describes the application of DEA by looking at data preparation from a SEMMA point of view, covered in Section 4.2 (Chapter 4), and followed by Section 5.4 analysing DEA results.

5.4 Data preparation and transformation

The focus of this chapter is to determine the efficiency at which a branch operates. The operating efficiency will be used in the following chapter, with supplementary data describing the external market environment, to predict the probability of a potential new branch operating efficiently. In addition, the results will also be used to highlight inefficiencies in the current branch network, the second objective of this study as described in Chapter one.

To ensure unbiased data envelopment results, CA, (Chapter 4), was committed to the identification of homogeneous branches. In the following section, variables selected in the previous section, describing the efficiency of branches, will be used to determine the operating efficiency of each homogeneous branch. Data used in DEA was provided by the financial institution and covered the period from 1 January 2007 to 31 December 2007. Microsoft Office Excel 2003 was used to implement the CRS (constant return to scale) version of the DEA model. The model was executed once for every branch. Subsequent sections cover data preparation prior to the application of DEA.

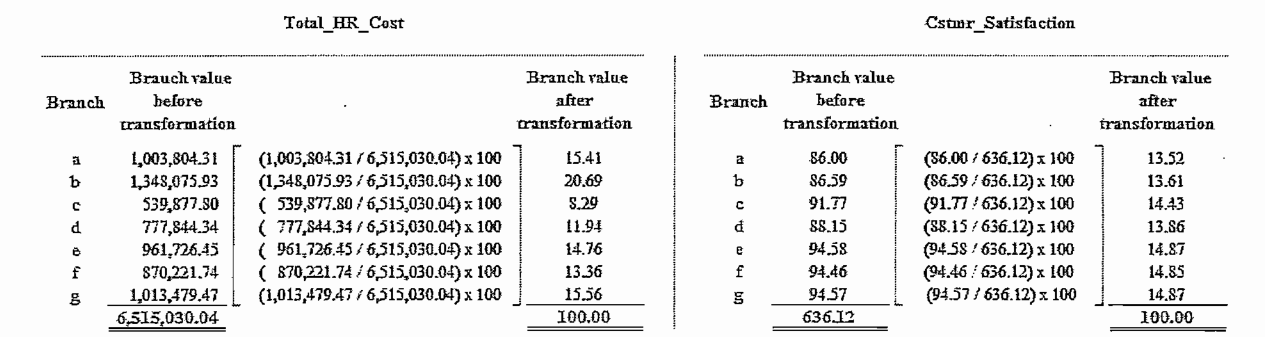
5.4.1 Exploring the data envelopment analysis data

CA identified 109 branches sharing a high level of similarity, which can be used for efficiency analysis. During data exploration, nine branches with incomplete and/or missing data surfaced and were excluded from analysis. Thus, from 109 similar branches initially identified, 100 branches will be used for DEA.

5.4.2 Modifying and transformation of the data

An advantage of DEA, according to Charnes *et al.* (1994:5), is that calculations can simultaneously utilise multiple outputs and inputs with each being stated in different units of measurement. Examples are human resource expense (Total_HR_Cost) and customer satisfaction (Cstmr_Satisfaction), as is the case in this study. However, as large differences in measurements pose a severe problem for Microsoft Office Excel 2003, it was necessary to scale variable values.

It was decided to scale variables to a value greater than zero and less than one hundred. Figure 5.2 displays, as an example, the transformation of seven branch values for human resource expense and customer satisfaction. In the example, branch A had a total human resource expense of R1 003 804.31 per month and a customer satisfaction value of 86%. After scaling, A had values of 15.41 and 13.52 for human resource expense and customer satisfaction respectively. Scaling of variables resolves the potential problem of large differences in measurements for Microsoft Office Excel 2003. All variables values used in DEA were scaled accordingly.



Total_HR_Cost

Branch	Brauch value before transformation		Branch value after transformation
a	1,003,804.31	(1,003,804.31 / 6,515,030.04) x 100	15.41
b	1,348,075.93	(1,348,075.93 / 6,515,030.04) x 100	20.69
c	539,877.80	(539,877.80 / 6,515,030.04) x 100	8.29
d	777,844.34	(777,844.34 / 6,515,030.04) x 100	11.94
e	961,726.45	(961,726.45 / 6,515,030.04) x 100	14.76
f	870,221.74	(870,221.74 / 6,515,030.04) x 100	13.36
g	1,013,479.47	(1,013,479.47 / 6,515,030.04) x 100	15.56
	6,515,030.04		100.00

Cstmr_Satisfaction

Branch	Branch value before transformation		Branch value after transformation
a	86.00	(86.00 / 636.12) x 100	13.52
b	86.59	(86.59 / 636.12) x 100	13.61
c	91.77	(91.77 / 636.12) x 100	14.43
d	88.15	(88.15 / 636.12) x 100	13.86
e	94.58	(94.58 / 636.12) x 100	14.87
f	94.46	(94.46 / 636.12) x 100	14.85
g	94.57	(94.57 / 636.12) x 100	14.87
	636.12		100.00

Figure 5.2 Transformation of variable values for DEA.

This section covered data preparation from a DEA perspective. Branches lacking information were removed from the analysis. Furthermore, to reduce the magnitude in differences of scale, all variables were scaled to a value between zero and one hundred.

5.5 Evaluating branch efficiency by means of data envelopment analysis

The evaluation of branch efficiencies to be presented in the following section can be categorised as follows: Firstly, Section 5.5.1 will cover the 100 branches selected from the branch network. This overview provides valuable insight into the overall efficiency, or health of the branch network. This section will cover aspects such as the number of branches operating efficiently and the identification of the most influential branch. Efficient and inefficient branches will be compared by looking at the utilisation of inputs and outputs produced. The scaled data (Section 5.4.2) used as input into DEA is listed in Appendix B. Owing to the sensitive nature of the information, the scaled values are listed and not the actual values.

Although the analysis in Section 5.5.1 is informative, it does not indicate how to improve the efficiency of a specific inefficient branch. Therefore, Section 5.5.2 furnishes a detailed analysis of a branch (15), identified as operating inefficiently, for presentation to management. Akin analysis for every inefficient branch was done and results presented to management, whom expressed their content with findings presented.

5.5.1 Preliminary analysis of DEA results

The efficiency score obtained by solving a DEA linear programming problem per branch is listed in Table 5.3. Each row in the table represents a solution with the objective of maximising the respective outputs for inputs utilised.

Of the 100 branches, 23 were identified as efficient (E=1) and 77 as operating inefficiently (E<1). The inefficient values range from a maximum value of 99% to a minimum of 36% with a standard deviation of 0.194 and a mean of 79.68%.

Every inefficient branch (E<1) has a corresponding reference subset listed next to it. The reference subset consists of a branch, or branches, producing a greater level of output with fewer inputs. Inefficiency of a branch is stated as compared to the reference branch, or set of branches. For example, a branch (15) is 88% efficient when compared to its reference set (branches 352; 422; 583 and 795) while branch 937 is 36% efficient when compared to reference set (branches 303 and 932).

The efficiency rating indicates in proportional terms, the maximum reduction across all inputs without causing a reduction in the output of the branch. Hence, the efficiency rating of 88% for branch 15 broadly implies that branch 15 can reduce its total inputs by 12% while still maintaining the same level of output. A more detailed analysis (presented in the following section) will provide specific recommendations pertaining to inputs for branch 15.

Br_Num	Efficiency score (E)	Reference set (reference branch = reference value)					
15	0.88	352 = 0.362	422 = 0.529	583 = 0.230	795 = 0.07		
20	0.84	63 = 0.911	795 = 0.023				
45	0.44	552 = 0.376	583 = 0.272	743 = 0.095	932 = 0.306		
49	0.52	63 = 0.618	932 = 0.290				
62	0.86	63 = 0.914	932 = 0.054				
63	1						
79	0.51	303 = 0.147	422 = 0.376	552 = 0.153	583 = 0.314	932 = 0.101	
83	1						
84	0.43	63 = 0.065	303 = 0.069	552 = 0.728	743 = 0.009	932 = 0.222	
88	0.81	93 = 0.356	352 = 0.624				
90	0.74	63 = 0.287	93 = 0.211	303 = 0.091	743 = 0.041	932 = 0.344	
92	0.83	93 = 0.095	743 = 0.227	932 = 0.687			
93	1						
124	0.73	63 = 0.062	303 = 0.174	552 = 0.803	743 = 0.004	932 = 0.104	
148	0.55	63 = 0.566	229 = 0.028	344 = 0.064	422 = 0.345	932 = 0.04	
185	0.87	63 = 0.219	795 = 0.341	932 = 0.381			
203	0.59	583 = 0.072	743 = 1.518				
214	0.92	93 = 0.031	743 = 0.489	932 = 0.502			
216	0.92	352 = 0.845	422 = 0.630				
221	0.57	422 = 0.175	768 = 0.226	795 = 0.273	932 = 0.267		
229	1						
238	0.73	583 = 0.282	743 = 0.505	932 = 0.184			
243	0.58	422 = 0.165	768 = 0.593	932 = 0.222			
251	0.45	63 = 0.667	932 = 0.229				
289	0.92	63 = 0.557	538 = 0.076	795 = 0.109	932 = 0.226		
303	1						
313	1						
314	0.78	93 = 0.020	303 = 0.812	743 = 0.145			
315	0.97	63 = 0.362	932 = 0.673				
318	0.96	63 = 0.313	795 = 0.618	932 = 0.093			
321	0.44	63 = 0.365	932 = 0.515				
323	0.95	313 = 0.242	352 = 0.228	422 = 0.453	671 = 0.097	795 = 0.304	
335	0.97	93 = 0.090	303 = 0.145	313 = 0.019	352 = 0.206	422 = 0.537	932 = 0.091
344	1						
346	0.91	303 = 0.058	313 = 0.249	422 = 0.243	583 = 0.099	743 = 0.548	
349	0.99	303 = 0.057	422 = 0.626	427 = 0.534			
352	1						
382	0.82	303 = 0.327	583 = 0.514	743 = 0.24			
397	0.6	63 = 0.083	93 = 0.522	422 = 0.347	486 = 0.004	538 = 0.039	
404	0.96	63 = 0.095	422 = 0.255	583 = 0.355	743 = 0.168	795 = 0.033	932 = 0.126
406	1						
415	0.76	63 = 0.439	422 = 0.198	743 = 0.328	795 = 0.051		
419	0.88	303 = 0.365	313 = 0.063	743 = 0.848			
422	1						
427	1						
428	0.62	313 = 0.285	422 = 0.304	486 = 0.232	538 = 0.006	743 = 0.091	
432	1						
440	0.44	63 = 0.477	932 = 0.496				
445	0.7	313 = 0.031	352 = 0.120	422 = 0.264	583 = 0.178	795 = 0.053	932 = 0.219
449	0.7	63 = 0.303	422 = 0.080	932 = 0.611			
466	0.93	63 = 0.176	313 = 0.061	422 = 0.494	538 = 0.253	743 = 0.039	795 = 0.062
486	1						

Br_Num	Efficiency score (E)	Reference set (reference branch = reference value)					
490	0.91	63 = 0.344	229 = 0.049	422 = 0.037	795 = 0.599		
509	0.94	303 = 0.135	422 = 1.764				
514	0.73	303 = 0.473	583 = 0.59	743 = 0.002			
527	0.66	303 = 0.349	422 = 0.014	583 = 0.061	932 = 0.639		
532	0.47	63 = 0.732	932 = 0.202				
533	0.77	422 = 0.097	768 = 0.355	795 = 0.514	932 = 0.053		
535	0.91	229 = 0.303	422 = 0.648	795 = 0.032			
536	0.52	63 = 0.487	932 = 0.424				
537	0.75	93 = 0.05	303 = 0.135	422 = 0.774	486 = 0.246		
538	1						
546	0.43	303 = 0.132	422 = 0.107	552 = 0.577	932 = 0.267		
550	0.52	63 = 0.258	932 = 0.783				
552	1						
572	0.6	422 = 0.095	552 = 0.195	583 = 0.183	743 = 0.104	932 = 0.471	
574	0.83	63 = 0.683	93 = 0.064	538 = 0.183	743 = 0.012	932 = 0.027	
580	0.88	352 = 0.139	422 = 1.328				
583	1						
596	0.45	63 = 0.176	932 = 0.738				
602	0.77	93 = 1.185	352 = 0.295				
670	0.49	63 = 0.670	932 = 0.299				
671	1						
689	0.57	538 = 0.017	743 = 0.766	795 = 0.094	932 = 0.145		
692	0.97	63 = 0.117	422 = 0.241	795 = 0.37	932 = 0.305		
717	0.95	63 = 0.218	422 = 0.164	932 = 0.652			
729	1	93 = 0.704	313 = 0.150	352 = 0.436			
735	0.61	303 = 0.476	552 = 0.596	932 = 0.04			
739	0.62	93 = 0.077	743 = 0.562	932 = 0.424			
743	1						
747	0.8	303 = 0.142	313 = 0.371	422 = 0.030	427 = 0.239	932 = 0.210	
757	0.87	303 = 0.142	486 = 0.327	552 = 0.210	743 = 0.378		
768	1						
772	0.88	63 = 0.832	795 = 0.046	932 = 0.053			
783	0.71	63 = 0.055	303 = 0.611	313 = 0.109	422 = 0.059	743 = 0.024	932 = 0.151
784	1						
795	1						
796	0.89	63 = 0.583	313 = 0.028	422 = 0.001	743 = 0.058	795 = 0.083	932 = 0.185
808	0.58	352 = 0.095	422 = 0.536	671 = 0.252	795 = 0.251		
835	0.9	63 = 0.516	932 = 0.469				
849	0.68	63 = 0.112	303 = 0.078	406 = 0.225	422 = 0.145	552 = 0.353	743 = 0.172
853	0.85	303 = 0.791	422 = 0.078	427 = 0.040	583 = 0.22		
862	0.6	303 = 0.173	422 = 0.447	552 = 0.192	932 = 0.178		
877	0.98	63 = 0.321	422 = 0.508	486 = 0.003	538 = 0.091		
884	0.5	63 = 0.410	932 = 0.593				
899	0.84	63 = 0.097	93 = 0.281	538 = 0.018	743 = 0.625	932 = 0.019	
932	1						
937	0.36	303 = 0.179	932 = 0.897				
939	0.84	352 = 0.552	795 = 0.492				
980	1						

Table 5.3 Branch efficiency ratings calculated by means of DEA.

A high level overview of the branch network operating efficiency is gained through the construction of a histogram. Figure 5.3 displays the histogram result of the efficiency ratings spread for the branch network.

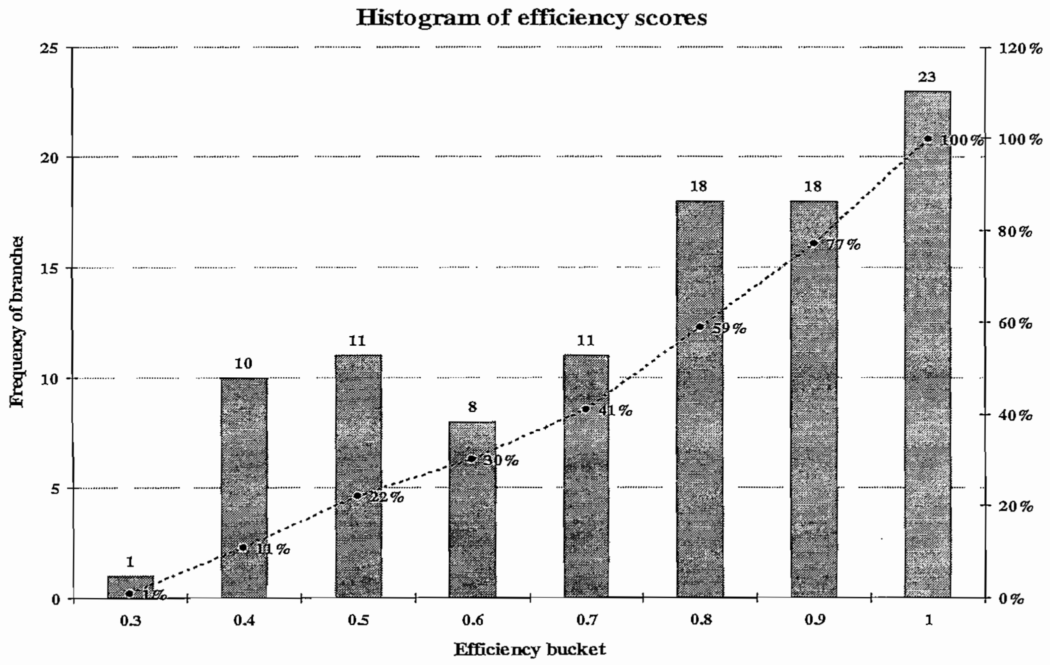


Figure 5.3 Histogram showing the efficiency spread.

According to efficiency measured, 23% of the branches operate efficiently (E=1), while one branch operates at an efficiency level of less than 40%. Management was perturbed with this information while expecting, in general terms, more branches to operate efficiently. This prompted further investigation into branch efficiencies. The fact that branches evaluated, share a high level of homogeneity, amplified management's concern, as results therefore cannot be attributed to economy of scale.

DEA does not rank branches in order of efficiency; consequently, it is not possible, solely by looking at efficiency scores, to determine which one of the 23 branches is the most prominent branch. Identification of this most influential branch is done through frequency analysis. The frequency of occurrence of an efficient branch is determined in the various reference subsets. Figure 5.4 shows the reference frequency analysis of the efficient branches.

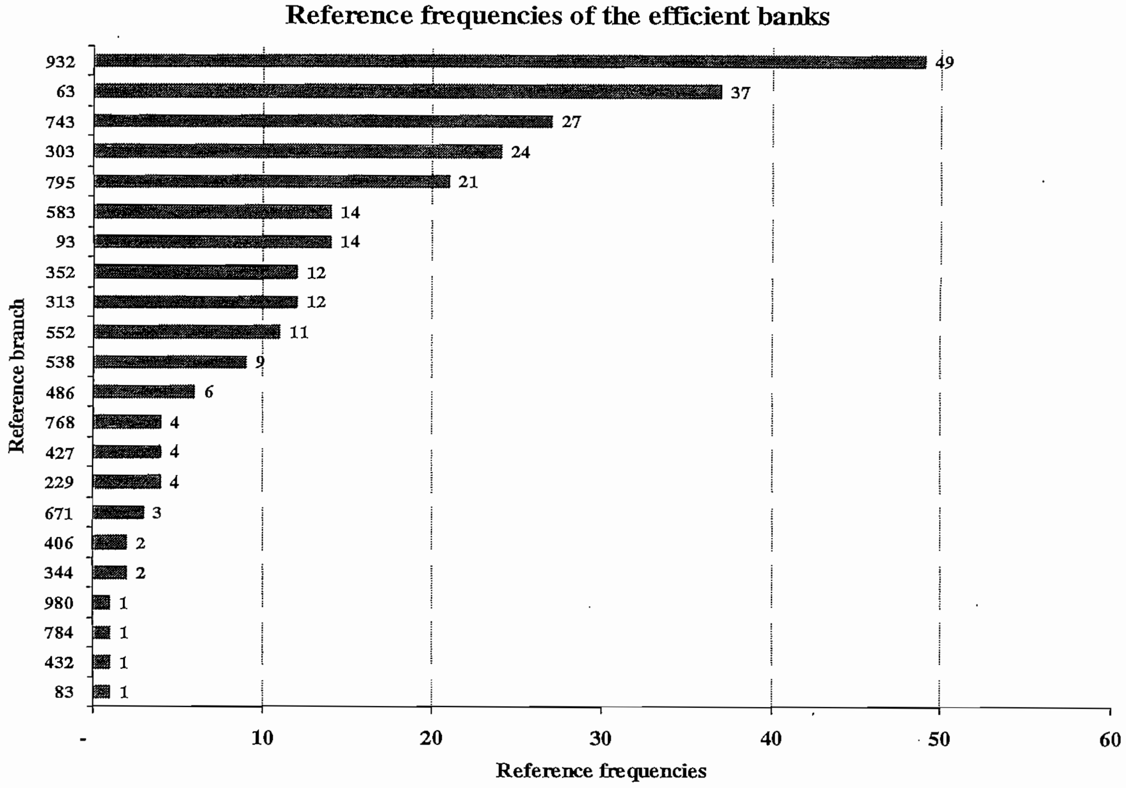


Figure 5.4 Reference frequencies for the efficient branches.

Mostafa (2008:316) states that branches with a high frequency count often appearing in reference sets of inefficient branches operate efficiently with regard to a large number of factors. Therefore, analysing such a branch is ideal for understanding the essential characteristics of efficiency.

On the other hand, branches with a low frequency count consist of an odd input / output mixture, therefore making such a branch, even when efficient, not suitable for explanatory analysis. From Figure 5.4, it is evident that branch 932 (49 times) and branch 63 (37 times) are highly affluent branches. Branches 83, 432, 784 and 980 with reference frequencies of one, although efficient, should be treated with additional care, should these branches be used for explanatory purposes.

An evaluation of efficient branches compared to inefficient branches is listed in Table 5.4. This evaluation directly compares the average outputs and inputs of efficient branches against that of inefficient branches.

		Efficient branch averages per output and input	%Difference	Inefficient branch averages per output and input
Outputs	Customer satisfaction (Br_Cstmr_Satisfaction)	0.986	⬇ -2%	1.004
	Transactional volume (Total_Txn_Volume)	1.134	⬆ 15%	0.960
	Sales volume (Total_Sales_Volume)	1.206	⬆ 22%	0.938
	Net interest income (Total_NII)	1.396	⬆ 37%	0.882
Inputs	Human resource cost (Total_Human_Cost)	0.808	⬇ 31%	1.057
	Rental cost (Total_Rental_Cost)	0.827	⬇ 27%	1.052
	Operational cost (Total_Operational_Cost)	0.891	⬇ 16%	1.033

Table 5.4 Detailed evaluations of efficient branches compared to inefficient branches.

The comparison result presents key characteristics making a distinction between efficient and inefficient branches. Efficient branches generate 15% more transactions, have a 22% higher sales volume and a net interest income of 37% more than their inefficient counter parts. It was interesting to note that efficient branches had a 2% lower customer satisfaction rating. While the efficient branches were producing significantly higher levels of output, they utilised 31% less human resources, 27% less rental cost and on average, had an operating cost of less than 16%.

The analysis presented in Table 5.4 highlights key aspects for management's attention. Although management was aware of the importance of managing human resources appropriately, Table 5.4 shows that human resource expense (Total_Human_Cost) has a major influence on branch efficiency, followed closely by rental cost (Total_Rental_Cost). Net interest income (Total_NII), with 37%, appeared to be the output characteristic efficient branches managed best.

Customer satisfaction is a very important aspect in the financial industry, and financial institutions spend a great deal of money, using various methods, to accurately portray its customers' sentiment. Results in Table 5.4 display efficient branches with a 2% lower customer satisfaction rating (Br_Cstmr_Satisfaction) than inefficient branches. This did not surprise management, as for some time, they suspected the inadequacy of current evaluation methods for determining customer satisfaction, but lacking verification of the fact. Results of this study confirmed management's opinion, and strengthened by the DEA results, initiated an investigation into customer satisfaction measuring methods.

This section analysed the current branch network as a holistic unit, in other words on a general basis comparing characteristics of efficient branches to inefficient branches. To fully understand the inadequacies at branch level, every branch should be scrutinised individually. Section 5.5.2 dissects branches individually in order to uncover inefficiencies associated with them.

5.5.2 Detailed analysis of individual branch efficiency

Beyond simply discovering inefficient branches, DEA provides additional insight relating to the magnitude of inefficiency associated with inputs and outputs. This ability of DEA makes it possible to pinpoint particular areas for improvement of inefficiency. Table 5.5 is an extract of Table 5.3 specifically referring to branch 15 and lists its efficiency score compared to that of the reference branches.

Br_Num	Efficiency score (E)	Reference set (reference branch = reference value)			
15	0.88	352 = 0.362	422 = 0.529	583 = 0.230	795 = 0.070
352	1				
422	1				
583	1				
795	1				

Table 5.5 Extract of DEA efficiency scores.

Every inefficient branch has a reference branch and a reference value. The reference value, also referred to as the shadow price, represents the amount of deficiency the inefficient branch has when compared to an efficient branch. Thus, branch 15, when compared to branch 352, had a deficiency of 0.362. The deficiency value indicates how much the efficiency of an inefficient branch (15) can be improved if inputs for the branch was reduced by one unit.

Data in Table 5.5 indicate that branch 15 operated at an efficiency level of 88% when compared to branches 352, 422, 583 and 795. This implies that the combination of efficient branches, 352, 422, 583 and 795, could achieve at least the same level of output as branch 15 while utilising 12% less inputs. Alternatively, branch 15 can reduce its inputs by 12% without experiencing any negative effects on output volume. Even though it has been established that branch 15 utilised more inputs than theoretically needed, it was still not clear which particular inputs were used excessively, as well as the magnitude of input abuse. To obtain insights regarding this fact, actual inputs and outputs of branches 15, 352, 422, 583 and 795 must be observed. Table 5.6 lists the actual units used.

	Input			Output			
Br_Num	Total_ Human_Cost	Total_ Rental_Cost	Total_ Operational_Cost	Br_Cstmr_ Satisfaction	Total_Txn_ Volume	Total_Sales_ Volume	Total_NII
15	1.496	0.512	1.041	0.963	1.256	1.59	1.949
352	1.390	0.685	1.064	0.961	1.242	1.647	4.562
422	0.604	0.209	0.583	0.915	0.775	1.584	0.272
583	0.723	0.357	0.756	1.011	1.411	0.611	0.663
795	0.519	0.133	0.675	0.961	1.03	0.219	0.026

Table 5.6 Actual data used by DEA.

Table 5.5 indicates branch 15 as an inefficient branch. However, it is difficult to determine the reasons for branch 15's inefficiency by merely looking at the information in Table 5.6. Total net interest income (Total_NII) and human resource expenditure (Total_Human_Cost) can be used to illustrate the point. Although branch 15 exploits more units of human resource expenditure, it achieved a total net interest income that is more than that of branches 422, 583 and 795 collectively.

To accurately establish areas of improvement, a composite value representing the efficiency reference set collectively, must be determined. The composite value is constructed by applying the reverence values of the reference branches to the actual inputs and outputs of the branch. Units used by the inefficient branch have to be compared to the composite value in order to reveal areas of improvement. Figure 5.5 displays the calculation of the composite values to which branch 15 should be compared for the purpose of highlighting inefficiencies.

		Shadow price for branch 352	Input / Output vector from branch 352	Shadow price for branch 422	Input / Output vector from branch 422	Shadow price for branch 583	Input / Output vector from branch 583	Shadow price for branch 795	Input / Output vector from branch 795		Composite value of efficiency reference set
Input	Total_Human_Cost		1.390		0.604		0.723		0.519		1.023
	Total_Rental_Cost		0.685		0.209		0.357		0.133		0.449
	Total_Operational_Cost		1.064		0.583		0.756		0.675		0.913
Output	Br_Cstmr_Satisfaction	0.36 X	0.961	0.53 X	0.915	0.23 X	1.011	0.07 X	0.961	=	1.131
	Total_Txn_Volume		1.242		0.775		1.411		1.030		1.255
	Total_Sales_Volume		1.647		1.584		0.611		0.219		1.588
	Total_NII		4.562		0.272		0.663		0.026		1.941

Figure 5.5 Calculation of the composite value.

Creation of the composite branch value makes efficiency analysis at branch level easy. Table 5.7 displays detail relating to the comparison of branch 15 to the composite value representing the efficient branch set. This table reflects the quantification of inefficiencies associated with branch 15.

	Variable	Branch 15	Composite efficiency reference set	Difference	% Difference
Input	Total_Human_Cost	1.496	1.023	0.47	46%
	Total_Rental_Cost	0.512	0.449	0.06	14%
	Total_Operational_Cost	1.041	0.913	0.13	14%
Output	Br_Cstmr_Satisfaction	0.963	1.131	-0.17	-15%
	Total_Txn_Volume	1.256	1.255	0	0%
	Total_Sales_Volume	1.59	1.588	0	0%
	Total_NII	1.949	1.941	0	0%

Table 5.7 Comparison between branch 15 and the reference branch set.

Results in Table 5.7 indicate that should branch 15 adopt the operating mixture of its reference branch set, the branch can reduce human resource expenditure (Total_Human_Cost) by 46%, rental expenses (Total_Rental_Cost) and operational expenses (Total_Operational_Cost) both by 14%, while still attaining the same level of output. Management was enlightened with the remarkable information provided by DEA. Besides the fact that branch 15 was operating inefficiently, management now also had a focus area for improving the operating efficiency of the branch. Detailed analysis similar to the above was performed for all inefficient branches and handed to management.

5.6 Chapter conclusion

This chapter was dedicated to separating efficient branches from inefficient branches. In addition, this chapter highlighted aspects causing inefficiency of a branch by comparing such inefficiency to an efficient branch reference set. The chapter started with 109 homogeneous branches identified in Chapter 4. Removing heterogeneous branches in the previous chapter, allowed for performing DEA without the risk of biasness in this chapter. Data exploration revealed that nine branches had inadequate information and were removed from analysis.

Data envelopment was applied to the remaining 100 branches, and the results revealed that twenty three branches operated efficiently. Benchmarking branch 15, an inefficient branch, against its efficient branch set, revealed an over expenditure on human resources by this branch.

The model provided the basis for reviewing and evaluating branch operations, thereby revealing operational differences between inefficient and best-practice branches.

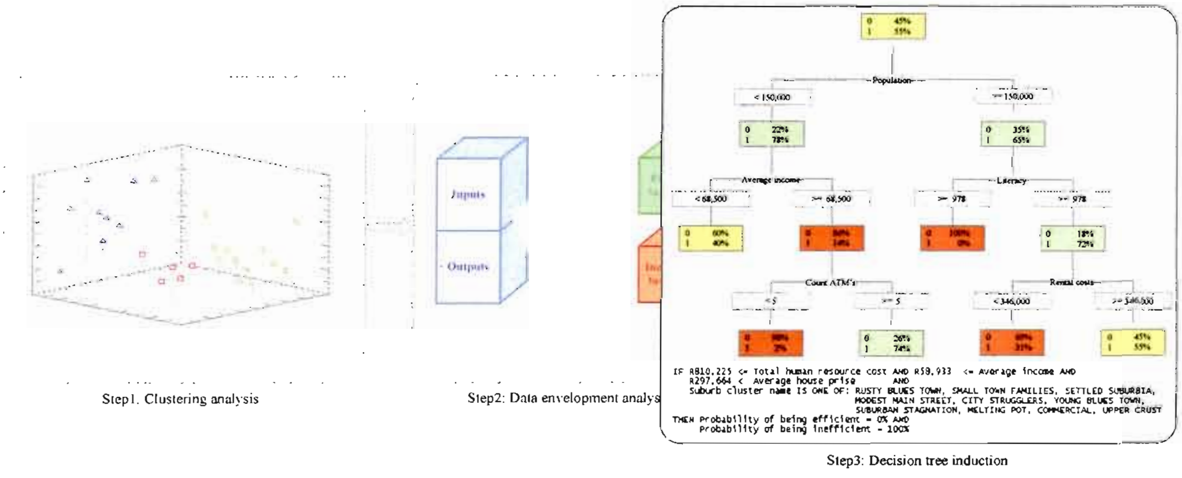
The next chapter will explain branch efficiency. Branches evaluated as efficient (E=1) will be assigned an efficiency label of 1, while branches with an efficiency of E<1 will be labelled 0 as being inefficient. The cluster membership will be used as the target variable of the decision tree (DT), while environmental variables representing the market environment the branch operates in, will be used as explanatory variables. DTI, the last of a three step process, will be discussed.

Chapter 6 - Decision tree induction to extract knowledge from data

6.1 Introduction

Decision tree induction (DTI) is the third and final step in the analytical process (Figure 6.1). The importance of this phase is perhaps best described with a metaphor relating to the mining industry. Mining for precious, but meagre minerals starts with a search for the precious minerals, that is finding a location where these minerals are in abundance, and this represents clustering analysis (CA). Subsequent to the identification of a possible mineral source, a mass of loam needs to be processed in order to discern between valuable and waste materials, this representing data envelopment analysis (DEA). Finally, after lots of tough work, valuable raw materials combined with other materials, in the right quantities, are transformed into usable products as represented by the decision tree (DT) in the analytical process.

The first two steps of the analytical process, CA and DEA, were concerned with identification of branches and relevant branch information, a task resembling the initial mining process. The last step in the process utilises distilled data from the previous two phases, combines it with geo-demographical data and transforms raw data into knowledge. This knowledge may be seen as data driven knowledge.



```
IF R810,225 <= Total human resource cost AND R58,933  <= Average income AND
   R297,664 <  Average house prise       AND
   Suburb cluster name IS ONE OF: RUSTY BLUES TOWN, SMALL TOWN FAMILIES, SETTLED SUBURBIA,
                                  MODEST MAIN STREET, CITY STRUGGLERS, YOUNG BLUES TOWN,
                                  SUBURBAN STAGNATION, MELTING POT, COMMERCIAL, UPPER CRUST
THEN Probability of being efficient = 0% AND
     Probability of being inefficient - 100%
```

Figure 6.1 Graphical illustration of the third step of the analytical model.

The objective of the model is to establish which branches of the financial institution operate efficiently and then use DTI to explain the results. An additional advantage is to predict the probability of whether a proposed new branch will operate efficiently in a certain market environment.

The rationale of this model is to identify efficient branches of the current branch network. This was achieved through the use of internal data for performing CA and DEA. Internal data refer to data specific to the financial institution this study is based on. Examples of internal data would be total human resource expense, daily operational expense and number of accounts and customers. These were all the variables described and used during phase 1 and 2(Chapter 4 and Chapter 5). At this stage of the modelling process, 100 branches were identified as homogeneous of which 23 were identified as operating efficiently (Table 5.3).

Having acquired the knowledge about branch efficiency, further initiative would be to add additional descriptive data describing the market environment the branch is located in. This data does not specifically relate to the financial institution the study is based on and will be referred to as external data. Subsequently, a DTI model will be built to determine combinations of internal and external data resulting in branches operating efficiently, thus further explaining DEA results. The main advantage of DTs is that results can be converted easily into interpretable if-then rules, in particular for complex DTs (Lu & Chen, 2009:3538). Furthermore, the DTI results can be used to predict the probability of a potential new branch operating efficiently.

Since management is considering expansion of the branch network, a budgeted amount will be allocated to the new branch. The amount shall be adequate to cover branch instalment costs, which may vary depending on certain conditions. The newly added branch is confined to function within the pre-set budget and within a specific market environment. During the planning phase, management will have a good idea of funds available for daily operations, whilst also having several desirable locations for the branch in mind. Information relating to newly proposed branches may be used, as well as the if-then rules, to evaluate the likelihood of such a branch operating efficiently. Branches having a high probability of not being efficient can consequently be eliminated.

The following section discusses the data used for DTI, also covering a description of data initially identified for DTI. This section will deal with logic relating to data selected for DT modelling. Section 6.3 continues with data preparation and ends with data finally selected for the DTI. This is followed by Section 6.4 describing the DT model used. The last section elucidates results obtained from applying the DT model.

6.2 Data selected for decision tree induction

The previous chapters focused on the identification of comparable branches (Chapter 4 - CA) and the evaluation of the efficiency of such branches (Chapter 5 - DEA). This resulted in 100 branches grouped as comparable, with 23 of them marked as operating efficiently. These branches with specifically selected variables, all of which are classified as internal data, will be used as input to the DT model. In addition to internal data used in DEA, data external to the financial institution, geo-demographical data, will be used for DTI. The variables selected from the DEA phase as input, include human resource expense (Total_HR_Cost), rent expense (Total_Rent) and operational expense (Total_Operational_Cost).

The objective of DEA (Chapter 5) was to allocate an efficiency score of E=1 for branches operating efficiently, and $0 \leq E < 1$ for branches operating inefficiently. The efficiency variable calculated in DEA is exceptionally important as it will be used as the target variable, also known as the dependant variable, in DTI. The internal and external data together with the target variables are used to train the predictive model. The trained DT model can be applied to unseen data for making predictions.

6.2.1 Data external to the financial institution

Branches do not function in isolation, as the surrounding area, market environment, etc., impact directly on their efficiency (Cavell *et al.* 2002:69). Consequently, it is of the utmost importance that market environmental factors be taken into consideration when trying to further explain efficiency results obtained from DEA, or when planning to add a new branch to the network.

External data used for DTI was obtained from the analytical department of the financial institution investigated during the period 1 January 2007 to 31 December 2007. Advanced geographical methods linking a branch to a suburb to obtain fundamental suburban characteristics, were being used. For example, if a branch is located in a metropolitan area, that is an area with a population of more than 250,000, a 5km radius is used to calculate external factors for that specific branch. If a branch is located in an area with a population of less than 250,000, a 10km radius is used to determine the external factors relevant to its situation. Included in the external data is a suburb group classification, modelled by a leading South African marketing company known as The Knowledge factory. Suburb classification is modelled mainly on 2001 deeds and census data covering the entire South African population. Table 6.1 tabulates the internal and external data used for the study.

Internal data, relating specifically to the financial institution.	Human resource expense (Total_HR_Cost)
	Operational expense (Total_Operational_Cost)
	Rent expense (Total_Rent)
Data external to the financial institution describing the market environment	Number of other competitive banks (Cnt_Banks)
	Number of automated teller machines (Cnt_ATM)
	Average proxy income per household (Avg_Income)
	Average selling price of houses in suburb (Avg_HSE_Price)
	Unemployment level of population (UnEmployment_Pop)
	Literacy of population (Litracy_Pop)
	Suburb group classification (Sub_Group)
Target variable	Branch efficiency (Efficiency)

Table 6.1 Variables initially identified as potential input parameters for the DT model.

To get a sense of the level of competition, the number of competitive banks (Cnt_Banks) and automated teller machines (Cnt_ATM) were selected as input data for DT. Inclusion of these could possibly explain the impact of competitive banks on the branch's efficiency.

The standard of living in the vicinity of a branch is expressed by the average income (Avg_Income), average house price (Avg_HSE_Price), unemployment rate of the population (UnEmployment_Pop) and literacy of the population (Litracy_Pop). Average income is a proxy amount describing the income of residents in the area a branch is located, in relation to the rest of the country. The actual selling prices of houses in the neighbourhood of the branch were used to determine an average house price for the area. Inclusion of the average income and average house price as potential variables were considered feasible as affluent customers have very different financial needs compared to less affluent customers requiring basic needs. Such variables therefore serve to distinguish between these customers groups.

Branches in the current branch network are located in diverse suburbs, the residents of some being more affluent then those in others. In an attempt to separate the disadvantaged from the rest, unemployment and literacy of the population in the vicinity of a branch were selected as potential variables. Literacy of the population is the number of residents possessing only the basic ability to read and write. Unemployment represents the portion of the population willing and able to work, but unable to find a job. Both literacy and unemployment are based on census data.

The ten suburb group classifications (Sub_Group) are listed and described in Table 6.2. Cluster classification is computed using variables describing a suburb holistically and uses the following characteristics to group and cluster similar suburbs (Knowledge factory, 2005:3):

- socio-economic rank: income, property value, education and occupation;
- life stage: age, household and family structure;
- residence type: size, type and age of structure.

The next section covers data preparation, according to the SEMMA model, and final selection of variables for DTI.

6.3 Data preparation for Decision tree induction

The previous section discussed and motivated the information deemed relevant to describe branch efficiency using DTI. This section will deal with the process of data preparation, with data being explored and, if required, transformed in such a way as to allow DT to interpret the data meaningful.

6.3.1 Sampling the data for decision tree induction

When analysing hundreds of thousands of records, it is advisable to extract representative samples from the population to simplify the analysis and modelling process. In the case of this model, with a 100 branches available for analysis, no sampling of records will be done as it is a relatively small data set.

6.3.2 Data exploring for decision tree induction

Correlation analysis was applied to the ten numerical variables (Table 6.1) with the aim to identify and remove redundant variables. Suburb group classification (Sub_Group) was excluded from the correlation analysis, as it was a non numeric categorical variable. The same method was followed as in Chapter 4, when correlation analysis was performed for the same reason. The squared correlation coefficient is listed in Table 6.3.

Suburb group classification (Sub_Group)	Suburb group description
(A) SILVER SPOONS	The silver spoons comprise of the most exclusive neighbourhoods, inhabited by the elite of South Africa and their living standards are among the highest in the world.
(B) UPPER MIDDLE CLASS	Do not quite live up to the standard of silver spoons but enjoy a favourable lifestyle. They are well educated white-collar workers in the trade, commerce and financial industries.
(C) MIDDLE SUBURBIA	A noticeable difference in living standards differentiates the middle suburbia from the upper middle class, with this group consisting of middle and lower-middle income earners. The residents are reasonable educated with either secondary or matriculation qualification.
(D) COMMUNITY NESTS	Large blocks of flats and townhouse complexes primarily define the community nests suburbs. These suburbs are mostly close to the city centre and most residents rent their accommodation.
(E) LABOUR POOL	The labour pool comprises a mixture of dwelling types with houses in the majority. With less than half of the residents not having a matriculation certificate, they do not expect to be paid well.
(F) NEW BONDS	The first generation of new parents in the democratic South Africa. Highly important to this group is the right to own property. Unfortunately, accompanying this right is a heavy financial burden.
(G) TOWNSHIP LIVING	Distinctive of the township living group is low-cost accommodation with approximately two thirds of the residents not having matriculation certification and suffering from high unemployment ratios.
(H) TOWERING DENSITY	Towering density represents crude tenement blocks mostly overcrowded and on the brink of social collapse. These tenants are poorly educated, with generally low incomes and mostly engaged in blue-collar occupations.
(I) DIRE STRAITS	Dire straits represent townships characterised by overcrowded places and an infrastructure that has collapsed. Low education, high crime and low incomes rank this group among the poorest of South Africa.
(J) BELOW THE BREADLINE	With little to no infrastructure, these clusters developed with no form of planning. With incomes and education at a desperate low, community leaders fight an anxious battle to keep social ills, such as crime, in check.

Table 6.2 Suburb group definitions (Knowledge factory, 2005:3).

Similar to Chapter 4, a threshold value of $r^2 = 80\%$ was used. The correlation analysis indicated that number of other competitive banks (Cnt_Banks) and automated teller machines (Cnt_ATM) had a high correlation value of $r^2 \approx 0.97$, or 97%.

Variable name	Total_HR_Cost	Total_Operational_Cost	Total_Rent	UnEmployment_Pop	Cnt_Banks	Cnt_ATM	Litracy_Pop	Avg_Income	Avg_HSE_Price	Efficency
Total_HR_Cost	1									
Total_Operational_Cost	0.56	1								
Total_Rent	0.18	0.26	1							
UnEmployment_Pop	0.00	0.00	0.00	1						
Cnt_Banks	0.00	0.02	0.05	0.00	1					
Cnt_ATM	0.01	0.03	0.05	0.00	0.97	1				
Litracy_Pop	0.01	0.01	0.01	0.59	0.04	0.06	1			
Avg_Income	0.00	0.03	0.08	0.00	0.03	0.03	0.00	1		
Avg_HSE_Price	0.02	0.04	0.01	0.03	0.02	0.03	0.04	0.17	1	
Efficency	0.09	0.03	0.01	0.00	0.06	0.06	0.00	0.00	0.00	1

Table 6.3 Squared correlation coefficients.

This correlation is expected as branches always have a number of automated teller machines placed at the branch entrance. Thus, since number of branches correlate better with the target variable, efficiency, it was decided to remove the variable number of automated teller machines from further analysis.

6.3.3 Modification and transformation of the data

Decision trees are renowned for their capacity to handle noisy and missing data (discussed in Chapter 3). Therefore, no data modifications with regard to these problems were performed. In fact, changes such as binning of variable values into equal size bins, also known as a dimension reduction, may decrease the predictive power of DTs (Walsh, 2005:2-29).

DEA resulted in an efficiency score of $0 \leq E \leq 1$. Classification trees predict categorical target variables. Consequently, it was required to transform the efficiency score obtained from DEA into a categorical target variable. The efficiency for all branches with an efficiency score of $0 \leq E < 1$ was transformed to 0. Thus, an efficiency score of E=1 indicates an efficient branch, while an efficiency of $E = 0$ indicates an inefficient branch. The remainder of the variables did not require data transformation. DT algorithms implicitly interpret the numerical variables and create numerical ranges similar to categories of character variables. Table 6.4 lists the input variables used for DTI.

Internal data	Human resource expense (Total_HR_Cost)
	Rent expense (Total_Rent)
	Operational expense (Total_Operational_Cost)
External data	Number of banks (Cnt_Banks)
	Average income (Avg_Income)
	Literacy of population (Litracy_Pop)
	Unemployment population (UnEmployment_Pop)
	Average house price (Avg_HSE_Price)
	Suburb group classification (Sub_Group)
Target variable	Branch efficiency (Efficiency)

Table 6.4 Variables selected for DT modelling.

Three variables describing a branch from an internal perspective and six variables describing it from an external point of view were selected for DTI in an attempt to clarify and explain branch efficiency scores.

6.4 Decision tree modelling

DT modelling was done through the use of the SAS 5.2 Enterprise miner software package. The software allows the user to select the splitting criteria, pruning rules and to control parameters, such as minimum node size and maximum tree depth. This enables the user to replicate and customise most of the popular algorithms. Figure 6.2 shows the process flow for the DT modelled for this analytical process. The scaled data used as input for DTI is listed in Appendix C.

The process flows from left to right and starts with an input dataset (WORK.BDF_1), a data partition node, two DT models using different splitting criteria methods, Entropy reduction and Gini reduction, and ends with a model assessment node.

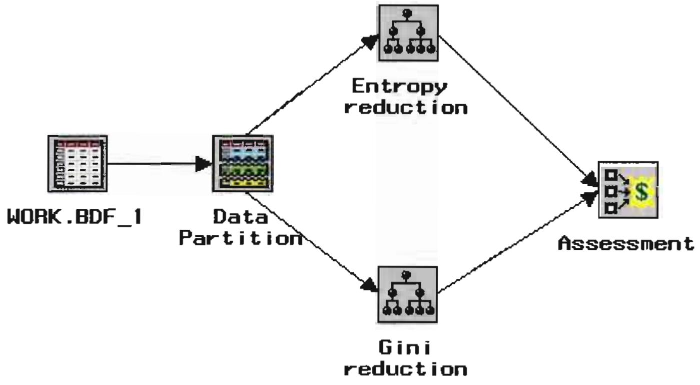


Figure 6.2 DT modelling with SAS Enterprise miner.

The input dataset contains the variables listed in Table 6.4, describing the 100 branches. For SAS Enterprise miner to determine how to use variables, the user defines the input dataset by specifying the variable model role and the variable measurement. Table 6.5 represents the input dataset definition.

Name	Model Role	Measurement	Type
BR_NUM	Id	Interval	num
TOTAL_HR_COST	Input	Interval	num
TOTAL_RENT	Input	Interval	num
TOTAL_OPERATIONAL_COST	Input	Interval	num
CNT_BANKS	Input	Interval	num
AVG_INCOME	Input	Interval	num
LITRACY_POP	Input	Interval	num
UNEMPLOYMENT_POP	Input	Interval	num
AVG_HSE_PRICE	Input	Interval	num
SUB_CLGROUP	Input	Ordinal	char
EFFICIENCY	Target	Binary	num

Table 6.5 Variables in the input dataset.

Branch number (BR_NUM) is a unique branch number and is assigned a model role of ID (identification). Assigning a variable a model role of ID, makes it possible to trace the path of a unique branch through the DT model. Important to note is the fact that the ID variable will not be used for modelling purposes. The role of target is assigned to the variable efficiency, indicating that efficiency is the variable the DT will predict. All the remaining variables are assigned the role of input, indicating that all variables are available for use to the DT algorithm.

Variable measurement for all variables, except for efficiency and suburb group classification (Sub_Group), is set to interval, as these variables are numerical and contain more than 10 distinct values. A binary model measurement is assigned to efficiency, indicating the fact that this variable contains two distinct values, $E = 1$, for efficient branches, and $E = 0$, for branches operating inefficiently. Suburb group classification (Sub_Group) is assigned a model measurement of ordinal, indicating the variable consists of a discrete set of values with the presence of a logical ordering (for example, Sub_Group: A, B and C).

The second node in the process flow (Figure 6.2), the data partition node, is used to randomly separate the input dataset into a training dataset, containing 70% of the branches in the input dataset, and a validation dataset, containing the remaining 30% of the input dataset. The training dataset is used to train the DT model and the validation dataset is used to prune and validate the DT model.

DT models mainly differ in the technique used to select the splitting attribute according to Zhao *et al.* (2009:2638). For the purposes of this model, two DT models were trained. The first model uses entropy reduction to determine the splitting variable and the second model uses Gini reduction to calculate the splitting variable. Han and Kamber (2006:297) describe Entropy reduction as a measure of how disorganised a system is. The Gini index computes the impurity of a dataset and then determines a splitting attribute.

The basic DT modelling settings, such as the minimum number of branches in a leaf and depth of the tree, etc. for both models were set to identical values. Table 6.6 displays the basic DT settings used for training both models.

Basic decision tree options	Value
Minimum number of observations in a leaf	1
Observations required for a split search	5
Maximum number of branches from a node	2
Maximum depth of tree	5
Splitting rules saved in each node	5
Surrogate rules saved in each node	0

Table 6.6 Basic options set for the DT model.

The growth of a DT can be limited by the use of various stopping rules. Two options used in this study is the minimum number of observations in a leaf and observations required for a split search. Sparse data available on branches required relatively small values assigned to the stopping rules (Table 6.6). For example, if the setting of minimum number of branches in a leaf were set to twenty, the tree would only result in a maximum of three splits, as it cannot grow with leafs having less than twenty branches.

The maximum number of branches from a node was assigned a value of two; this setting would allow only binary splits from nodes. DTs allowing multi-way splits, are not necessarily more powerful than DTs grown with binary splits and may result in exceedingly complex bushy trees. (Walsh, 2005:2-55). Therefore, to avoid an overly complex DT and aid in interpretability, only binary splits were considered for this application.

The DT modelling process ends with an assessment node. This node is used to compare different models and to select the model that best does the task of predicting. The following section will cover the DT results from two perspectives, firstly by evaluating the model built and secondly, by analysing the model results.

6.5 Interpreting decision tree results

DT results are discussed from two perspectives. Two DT models were built only differing in the method used to determine the splitting variables at a node. The first model uses Entropy reduction and the second model uses Gini reduction to calculate the splitting variable. Firstly, the model yielding the best results will be selected, using the model assessment node, and secondly, the results obtained by using the best model will be discussed.

The next section covers model selection where the DT yielding the best results will be selected, followed by Section 6.4.2 that will analyse and interpret the results of the DT selected as the superior tree.

6.5.1 Selecting the most effective prediction model

Lift charts are used to gain insight into the effectiveness of a model and aid in the selection of a model (D' Souza *et al.*, 2007:285). The Gini reduction and Entropy reduction DTs will be compared, measuring their prediction performance, using a cumulative percentage response (%Response) chart and percentage captured response (%Captured response) chart. The model performing best was selected to be used in this analytical model.

The %Response chart arranges branches into deciles based on their probability of operating efficiently and then plot the actual percentage of efficient branches (Walsh, 2005:2-55). A valuable model will have a relatively high value for predicted responses versus actual results. Figure 6.3 displays a %Response graph with two models depicted, the Entropy reduction based DT and the Gini reduction based DT. Also displayed in Figure 6.3 is the base line representing response results based on a random sample of the data.

Based on the %Response chart in Figure 6.3, the Gini reduction DT model performed better, with 74% of the response in the first two deciles as compared to 46% response in the first two deciles for the Entropy reduction DT model.

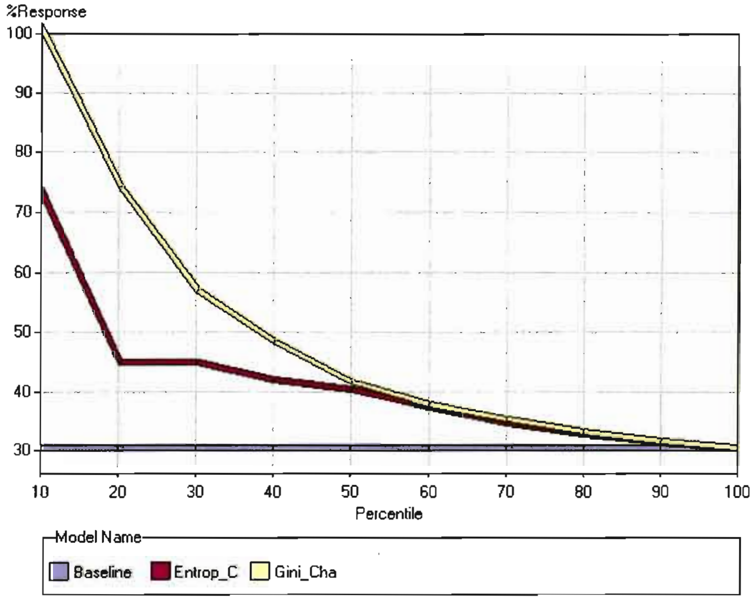


Figure 6.3 Cumulative %Response chart comparing the two models.

The %Captured response graph in Figure 6.4 displays the percentages of total number of responders in a decile. The Gini reduction DT performed better than the Entropy reduction tree model with a higher prediction rate in the first 50 deciles.

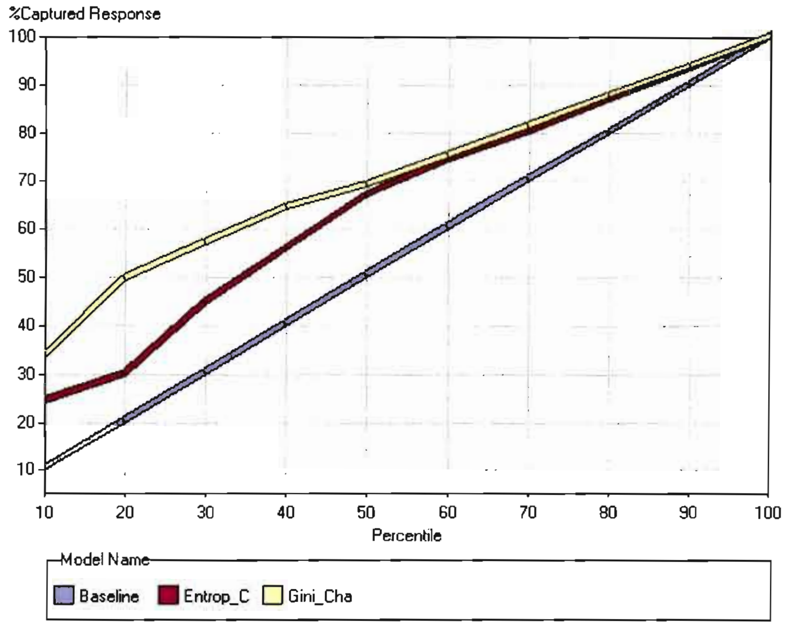


Figure 6.4 % Captured response graph comparison between the two models.

The %Response chart and the %Captured response chart indicate that for this specific application of DT, modelling the Gini reduction tree performed better than the Entropy reduction DT model and was thus selected as the preferred model for this analytical model.

6.5.2 Description of branch efficiency using decision tree results.

In the previous section, it was determined that the Gini reduction DT performed better than the Entropy reduction DT. It was therefore decided to use the Gini model. This section will analyse and interpret the results generated from the Gini DT. The section starts with a graphical representation of the Gini DT as depicted in Figure 6.5. The tree displays a classification tree with eight leaves.

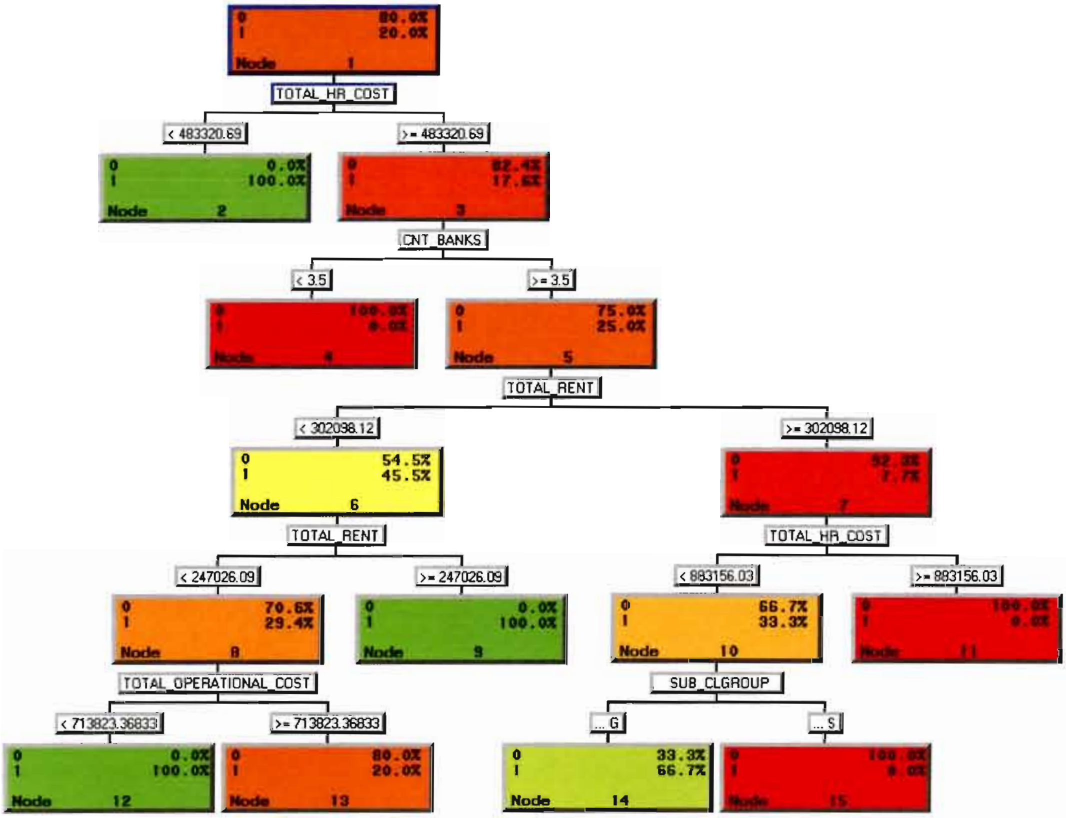


Figure 6.5 Graphical representation of the DT.

The following general inferences can be drawn from the tree diagram:

- Among the observed branches, twenty per cent operate efficiently; this is represented in the root node of the tree at the top of the diagram. The graphical display of the DT assists with model interpretation, with the aid of color coding, to identify leaves containing efficient branches. Leaves of the tree with a green background contain a high concentration of efficient branches, whereas leaves color coded with red, indicate a high concentration on inefficient branches;
- Human resource expense is a prime factor management should consider when evaluating data envelopment results, or when planning to open a new branch. According to the tree, branches seldom operate efficiently if the branch expenditure on human resources exceeds R483 320 per month;
- The number of other banks in the vicinity of the branch played a dominant role in the efficiency of a branch and represented the external factor playing the biggest part in describing branch efficiency;

- Unemployment population (UnEmployment_Pop), literacy of population (Litracy_Pop), average income (Avg_Income) and average house price (Avg_HSE_Price) had no influence on the creation of the DT. All of these variables describe the external market environment;
- Internal data, data relating specifically to the financial institution, were more significant in the tree creation. This fact might be an indication that management should severely revise general branch expenditure to increase branch operating efficiency.

The DT diagram aids in interpreting the DT in a graphical manner. The real power of DTs above methods like neural networks, which is a black box environment, is the fact that DTs translate into easy understandable rules. Table 6.7 represents the natural language version of the DT graphical diagram in terms easily understood.

Interpreting the top level of the DT diagram is relatively easy when considering rules one, two, three and four. Rule two indicates that, if there are less than three other banks in the vicinity of the branch and the branch's expenditure on human resources is greater than, or equal to R483320, it has a 100% probability of operating inefficiently.

Making conclusions of leaf nodes at the lower end of the DT diagram may be more difficult. For example, rules seven and eight derived from the DT diagram are much easier to interpret than analysing the diagram. These rules indicate that, if two similar branches - with regard to human resource expense, rent expense and the number of other banks in the area - were located in different suburbs, the efficiency of a branch might well be influenced. The efficiency of the branch will ultimately depend on the suburb the branch is located in.

Interestingly, when more than three branches of other financial institutions operate in the same vicinity, a branch from this financial institution has the propensity of operating efficiently. Management supported this statement, deducted from the DT, and elucidated that financial institutions usually select similar retail locations, based on the amount of potential sales offered, for instance shopping malls, etc. Results from the DT regarding this fact could be used to verify the fact, since prior to the results, no such facts supporting the view point, existed.

All the rules were given to management whom expressed their satisfaction with rules explaining branch efficiency. In analysing the results, they found the combinations of features resulting in a

100% operating inefficiency, as particularly informative. As a primary action, they will strive to identify instances where these features exist and attempt to eliminate them.

#	Condition	*N*	Efficiency status	Probability
1	Human resource expense < R 483,320	2	Efficient	100%
2	Number of banks < 3.5 Human resource expense >= R 483,320	20	Inefficient	100%
3	R 247,026 <= Rent expense < R 302,098 Number of banks < 3.5 Human resource expense >= R 483,320	5	Efficient	100%
4	Human resource expense >= R 883,156 Rent expense >= R 302,098 Number of banks < 3.5	20	Inefficient	100%
5	Operational cost < R 713,823 Rent expense < R 247,026 Number of banks >= 3.5 Human resource expense >= R 483,320	2	Efficient	100%
6	Operational cost >= R 713,823 Rent expense < R 247,026 Number of banks >= 3.5 Human resource expense >= R 483,320	15	Inefficient	80%
7	Suburb group classification IS ONE OF: A B C D E F G R 483,320 <= Human resource expense < R 883,156 Rent expense >= R 302,098 Number of banks >= 3.5	3	Efficient	66%
8	Suburb group classification IS ONE OF: H J S R 483,320 <= Human resource expense < R 883,156 Rent expense >= R 302,098 Number of banks >= 3.5	3	Inefficient	100%

Table 6.7 Pairs of rules describing relative branch efficiency.

With the use of rules, generated by DT, management may be able to identify conditions that lead to branches operating more efficiently. When planning to add new branches, they should evaluate the rules and exclude branches from consideration if a high probability of inefficiency is present.

6.6 Chapter conclusion

Results obtained from the model provided the basis for reviewing and evaluating branch operations. These findings revealed operational differences between inefficient and best-practice branches. DEA was used to determine operating efficiency of branches and a DT was constructed to use the efficiency results of DEA. Data describing the market environment was included in the DT with the aim to better understand circumstances leading to branch efficiency. In this method, the DT was used to enhance DEA results and optimise the value of information gained from DEA. The number of other financial institutions can be used as an example of additional value added to the data envelopment results. Without the inclusion of a DT, this fact would not have come to light.

The chapter started with variable selection and data preparation. The variable selection concluded with a final set of variables selected for DT modelling purposes. Two DTs with the exact modelling criteria were constructed. One model used Entropy reduction to determine the splitting variable at a node, whereas the other used Gini reduction to determine the variable that best separates the branches with respect to the target variable, efficiency.

The two DT models were compared using a %Response and %Captured response chart, with the Gini reduction model performing better at predicting efficient branches. The Chapter concluded with a discussion on the DT results and rules generated from the model.

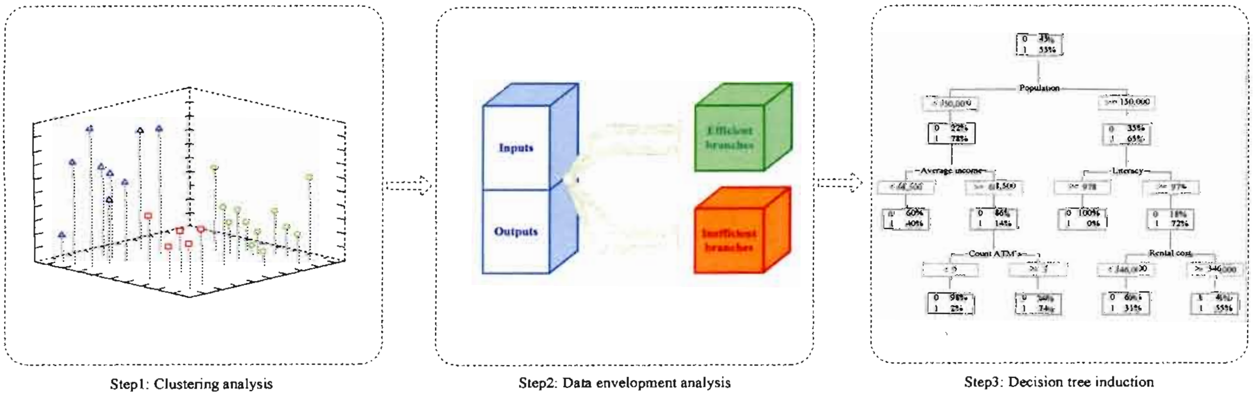
Chapter 7 - Summary

7.1 Introduction

The purpose of Chapter 7 is to present the final comments and concluding remarks on the research project. The objectives of the study and how they were achieved will be summarised. New problems and opportunities identified will also be outlined.

7.2 Overview of the research project

The aim of this study was to better equip managers of a financial institution's branch network in maintaining the network. Fulfilling the objectives empowered management to make better informed decisions, based on a sound analytical model, and improve the performance of branches, ultimately leading to a more profitable financial institution. In order to reach the objectives, the use of an analytical model making use of a combination of data mining and management science techniques, was proposed. Figure 7.1 displays the three step analytical model presented.



Step1: Clustering analysis

Step2: Data envelopment analysis

Step3: Decision tree induction

Figure 7.1 Graphical illustration of proposed analytical model.

The methodology employed may be summarised as follows: Chapter 1 raised a convincing argument as to why the study of branch network efficiency should receive high priority. Physical branches came under severe pressure from, seemingly, more cost effective alternative delivery channels. However, in a short span of time, it became evident that these alternative delivery channels would merely be seen as an additional delivery channel for financial institutions rather than replacing the branch. The realisation of the importance of branches may be seen as a revival of the branch network. By looking at the background of the branch network, one realizes the value of such a study.

Chapters 2 and 3 focused on underlying literature. A study was conducted to determine alternative research methodologies and to investigate the feasibility of such a study. Although regular contact sessions were held with representatives of the financial institution, no formal use was made of qualitative research methods, as the study was based on quantitative research methods. The large volume of literature available on the subject of efficiency measurement made it apparent that this was a matter receiving a great deal of attention. Literature references confirmed that an investigation into branch efficiency is achievable and that methods such as clustering analysis (CA), data envelopment analysis (DEA) and decision tree induction (DTI) were regularly employed to solving problems related to this subject.

The empirical part of the study was covered in Chapters 4, 5 and 6, with each chapter dedicated to the use of a specific analytical method. Chapter 4 utilised CA. In an attempt to improve clustering results, two rounds of CA were conducted. The intention of the first round was to remove outlier branches, whereas the second round focused on the creation of high quality clusters. The results of clustering results were presented to management. They confirmed that, based on expert domain knowledge, the clusters were indeed representative of similar branches. One can therefore conclude that the application of CA to identify homogeneous branches was successful.

Chapter 5 reported on the application of DEA, the second step in the three step process (Figure 7.1). DEA may be seen as the core of the analytical model. During DEA, homogeneous branches were compared to identify branches operating efficiently, thus meeting the second objective of the study. It provided management with clear information as to changes needed to improve the efficiency of inefficient branches.

To attain more knowledge from the DEA results, decision tree induction (DTI) with geo-demographic data describing the market environment, was used (Chapter 6).Motivation for the use of geo-demographic data to explain DEA results was the fact that branches do not function in isolation, as the surrounding area does impact significantly on branch efficiency. The results were presented to and confirmed by management. They were aware of the fact that branches from other financial institutions impacted directly on the efficiency of their branch network, but confirmation to this end was lacking. Results from DTI confirmed their view point.

7.3 Overview of research objectives

It was believed that, for management to improve the operating efficiency of the branch network, they would require further insight on factors impacting overall branch efficiency. The primary outcome of this study therefore was to create an analytical model that would amplify features influencing branch network performance. Results obtained from decision tree induction (Chapter 6), gave specific features required for branches to operate efficiently. Management obtained a better understanding of what drives branch operating efficiency and based on their domain knowledge, they were able to validate several of the findings.

The secondary objective of this study was aimed at branch maintenance by means of benchmarking individual branches within the branch network. Results obtained, directed management to specific features to be addressed for improving efficiency of an inefficient branch. Branch fifteen was used as such an example (Chapter 5).

A significant feature of the study was the fact that three distinct analytical methods from two different backgrounds were used to reach its objectives. The fusion of multiple techniques contributed to research and illustrated additional advantages to be gained from collaborating different analytical methods. In addition, data mining and management science techniques were demonstrated as having an important role to fulfil in the financial industry, especially since this industry focuses on the use of financial ratios rather than other analytical methods.

7.4 Limitations of the study

DEA was applied to one hundred of the homogeneous branches for determining branches that operate efficiently. The results revealed 23 branches operating efficiently, however DEA does not have the ability to identify a single branch as a best practise decision making unit. Therefore, questions regarding the best practise branch had to be answered using reference frequencies (Section 5.5.1, Chapter 5).

The DTI had 100 branches available for training and validating the DT. With such few records, the tree pruning rules, training and validation had to be kept to a minimum.

7.5 Future studies

The study was based on the branch network of one of the leading financial institutions in South Africa. Branches were compared to determine efficient branches and in doing so, gave managers the ability to better understand what is important for branch efficiency. As the study was limited to one financial institution only, results obtained might be biased towards specific methods employed by this concern. The implementation of a similar study across multiple financial institutions would eliminate this constraint.

The objective of the study was to assist management in managing the network of branches. Chapter 5 presented key features that could transform a branch from performing inefficiently to being efficient; branch 15 was used as an example in Chapter 5. Implementing a methodology that would allow management to measure the response time of changes, as well as the impact of such changes on the branch network, would represent a natural follow-up project.

7.6 Chapter conclusion

Chapter 7 is the final chapter of this study. The chapter presented a summary of the initial objectives and how they were achieved. In conclusion, possible future research opportunities were outlined.

Appendix A – Clustering analysis input

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
14	0.61414	0.48380	0.15393	0.05656	0.55370	0.30718	-0.16193	-0.46716	-0.56264	0.01485
15	-0.93479	-0.85659	-0.65433	-0.70636	-0.91283	-0.55990	-0.57541	-0.49476	-0.68891	-0.94760
20	-0.26954	-0.80751	-0.21556	-0.38638	-0.78438	-0.55990	-0.70068	-0.49476	-0.70469	-0.99990
21	-0.33905	-0.31823	-0.35229	-0.35005	-0.49937	-0.55990	-0.53565	-0.21881	-0.35746	0.57977
29	-0.43834	-0.10729	-0.14385	0.17324	-0.56359	-0.55990	0.38788	-0.46716	-0.68891	-0.29899
32	0.74322	0.83025	1.06731	0.34930	0.08805	0.30718	-0.04563	-0.10844	0.05290	1.07145
33	-0.01139	-0.09117	0.38182	0.15926	-0.52880	0.30718	-0.58558	-0.46716	-0.21541	0.10900
34	-0.44827	-0.29846	-0.50908	-0.37590	0.05861	-0.55990	0.07515	-0.30160	-0.34168	-0.45591
37	1.32903	1.09540	1.20101	0.08241	0.34764	0.30718	-0.26149	0.38826	1.36291	1.37483
39	0.18719	0.27212	0.49789	-0.33817	-0.03907	-0.55990	-0.48931	0.60902	1.20508	0.64253
44	1.14038	1.62863	2.84182	0.94734	1.64156	1.60780	0.31433	-0.16362	2.29412	1.65729
45	-0.53266	-0.71669	-0.52002	-0.71474	-0.82452	-0.55990	-0.80263	-0.43957	-0.24698	-0.44545
47	0.21202	0.55778	0.25055	-0.14884	0.63399	0.30718	-0.28093	0.16751	0.77893	0.97730
48	-0.28940	0.34171	0.03968	-0.42201	0.35433	1.17426	-0.33773	1.43685	0.79471	0.61115
49	-0.69649	-0.96646	-0.90409	-0.84958	-0.97840	-0.12636	-0.91325	-0.21881	-0.42060	-0.73837
57	0.06308	0.07509	-0.14263	0.19140	0.83738	-0.55990	0.76877	-0.32919	0.22652	0.11946
62	-0.45323	-0.66835	-0.78498	-0.69099	-0.03639	0.30718	-0.10782	-0.43957	-0.57843	-0.48730
63	-1.01422	-0.59144	-1.03597	-0.60436	-0.80445	-0.55990	-0.16492	-0.49476	-0.73626	-0.88483
78	-0.16529	0.26993	-0.03021	-0.44087	0.10812	-0.12636	-0.48632	0.80218	0.84206	0.46469
79	-1.08869	-0.93936	-0.66769	-0.34656	-0.97974	-0.55990	-0.59903	-0.49476	-0.76783	-0.94760
81	-0.64188	-0.41345	0.33320	0.40239	-0.42443	-0.55990	-0.08510	-0.46716	-0.46794	-0.43499
83	-1.02912	-0.91665	-1.18060	-0.97674	-1.07340	-0.55990	-0.96019	-0.49476	-0.76783	-0.85344
84	-0.24472	-0.75917	-0.78559	-0.64907	-0.85262	-0.55990	-0.80114	-0.19122	-0.49951	-0.68606

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
88	-0.89507	-0.92471	-0.87432	-0.83142	-0.91016	-0.55990	-0.86303	-0.41197	-0.38903	-0.59191
89	-0.19011	1.01483	0.53253	-0.35284	0.72631	-0.12636	-0.02172	0.27788	-0.38903	0.68438
90	-0.55749	-0.22302	-0.42218	-0.71405	-0.42711	0.30718	-0.73566	0.02954	-0.62578	0.30777
92	-0.64188	-0.77602	-0.78073	-0.69169	-0.87537	-0.55990	-0.77064	-0.46716	-0.70469	-0.34084
93	-0.54756	-0.42444	-0.48721	-0.38149	-0.35352	-0.55990	-0.63162	-0.30160	0.02134	-0.13161
101	-0.35891	0.24063	0.53921	0.77687	0.34898	-0.55990	0.18159	-0.35678	-0.12071	-0.07930
102	1.87017	2.24902	3.21131	4.33578	1.97207	-0.12636	2.59281	-0.16362	0.41592	0.55884
103	-0.96954	-0.97451	-0.65554	-0.61973	-0.87403	-0.55990	-0.78709	-0.49476	-0.62578	-0.78021
110	0.50989	0.53288	1.28062	-0.08806	0.00777	-0.55990	-0.31232	-0.05325	0.84206	0.87268
124	-0.63692	-0.54749	-0.14810	-0.63161	-0.52078	-0.12636	-0.72789	-0.49476	-0.54686	-0.60237
126	-1.01422	-1.15177	-1.16723	-0.92294	-0.91149	-0.55990	-0.74194	-0.49476	-0.76783	-0.78021
133	1.04606	0.98920	0.50275	0.11385	1.05013	0.74072	0.66054	0.58142	1.39448	1.30160
134	0.98648	0.44498	1.11168	0.66928	0.24862	0.30718	-0.12456	-0.49476	0.03712	-0.21530
137	1.16520	1.15546	0.83699	0.34021	0.82801	-0.55990	-0.09885	0.30548	-0.37325	0.70530
141	-0.60217	-0.04942	-0.04054	-0.01680	-0.27724	0.74072	-0.16044	-0.16362	-0.04180	-0.42453
144	0.91698	1.08808	0.58722	-0.13068	1.14380	2.04134	0.49461	1.57482	1.61544	0.94591
147	-0.46316	-0.76210	0.51794	0.94804	1.33648	-0.55990	0.03479	-0.46716	-0.65734	-0.85344
148	-0.56742	-0.71230	-0.80261	-0.74688	-0.84058	-0.55990	-0.71175	-0.35678	-0.56264	-0.60237
153	0.29641	0.40397	1.76679	1.46853	-0.23576	0.30718	0.56906	-0.49476	-0.37325	0.08808
159	0.45528	1.09540	1.90231	1.23937	1.26021	-0.12636	0.80524	-0.32919	0.08447	0.52746
172	-0.44827	-0.57459	-0.47019	0.01744	-0.01364	-0.55990	0.45156	-0.49476	-0.57843	-0.86391
174	-0.01635	0.52555	0.41038	-0.04614	0.53898	-0.55990	-0.22382	0.77458	0.71580	0.72622
176	2.02903	0.60100	0.36663	0.31716	-0.20633	0.30718	0.23779	-0.43957	0.81050	0.24500
184	-0.27451	-0.37317	0.70512	0.92847	-0.18224	-0.55990	-0.29797	-0.41197	-0.70469	0.03577
185	-0.59720	-0.48597	-0.57046	-0.74898	-0.57162	-0.55990	-0.74613	0.05713	-0.04180	0.07762
194	0.36095	1.27778	2.11379	1.27570	1.49303	-0.12636	0.57982	-0.30160	0.19495	0.21362
195	-0.62202	-0.58485	-0.68532	-0.16770	0.26468	-0.55990	0.36187	-0.43957	-0.48373	-0.75929
203	0.16733	-0.16881	-0.61604	-0.59318	-0.59437	-0.55990	-0.62654	-0.24641	-0.13650	0.15085

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
207	0.07301	-0.09484	-0.08065	-0.39686	-0.69473	-0.12636	-0.61488	-0.32919	0.27387	-0.05838
210	-0.30430	-0.22448	-0.35290	0.41008	0.16298	-0.55990	0.43900	-0.46716	-0.73626	-0.44545
211	-0.61210	0.00258	-0.04419	-0.36752	0.60187	0.30718	-0.31381	-0.10844	0.19495	-0.28853
213	1.34393	1.33711	0.98770	0.64203	1.65762	-0.55990	0.13046	0.00194	-0.43638	0.99822
214	-0.77593	-0.83681	-0.93326	-0.80557	-0.72149	-0.12636	-0.83104	-0.19122	-0.72048	-0.68606
215	0.05811	0.23404	0.17641	0.71329	0.53631	0.30718	-0.21904	0.27788	-0.68891	0.20316
216	-0.86032	-0.94815	-1.01591	-0.83561	-0.66930	-0.55990	-0.83672	-0.30160	-0.73626	-0.76975
217	-0.43834	-0.44422	-0.40395	0.47994	0.18038	-0.55990	0.80644	-0.46716	-0.68891	-0.62329
221	-0.33905	-0.54603	-0.13170	-0.57711	-0.64254	-0.55990	-0.68693	-0.43957	-0.59421	-0.36176
223	-1.11848	-1.19132	-1.17209	-0.95368	-0.99312	-0.55990	-0.94255	-0.30160	-0.76783	-0.99990
224	-0.82557	-0.82436	-0.87735	-0.88032	-0.83790	0.74072	-0.86153	1.35406	-0.38903	-0.65468
229	-0.60217	-0.86025	-0.88769	-0.64348	-0.76966	-0.55990	-0.75689	-0.49476	-0.75204	-0.40360
233	0.05315	0.19082	0.68932	1.01161	0.34764	-0.55990	0.53766	-0.43957	-0.76783	-0.18392
234	0.29145	0.44572	0.24873	-0.39476	1.02872	1.60780	0.15169	-0.16362	0.70001	0.14039
236	-0.85039	-0.80165	-0.37782	0.09708	0.65138	-0.55990	0.76189	-0.49476	-0.73626	-0.88483
238	-1.04897	-0.99502	-1.06088	-0.44227	-0.51676	-0.55990	0.02971	-0.49476	-0.75204	-0.94760
243	-0.40855	-0.81484	-0.07518	-0.69169	-0.75762	-0.55990	-0.71832	-0.49476	-0.73626	-0.92667
250	-0.70642	-0.07213	-0.15479	-0.20892	-0.13675	-0.55990	-0.37510	0.00194	0.16338	-0.07930
251	-0.59720	-0.80678	-0.65068	-0.79299	-0.56092	-0.55990	-0.67946	-0.49476	-0.70469	-0.92667
253	-0.03125	-0.23693	-0.32373	0.22633	-0.22105	-0.55990	-0.22472	-0.49476	-0.76783	-0.42453
254	-0.14543	-0.37976	-0.45621	0.19839	0.08537	-0.55990	-0.02829	-0.49476	-0.75204	-0.31991
255	-0.86529	-0.88808	-0.43190	0.85442	0.13355	-0.55990	0.63662	-0.49476	-0.72048	-0.57099
277	-0.19011	-0.28674	0.33928	0.37235	0.14157	-0.55990	0.22882	-0.43957	-0.64156	-0.66514
289	-0.94472	-1.01260	-1.05177	-0.90198	-0.77367	-0.55990	-0.82207	-0.46716	-0.51529	-0.87437
302	0.94180	0.30948	0.56048	0.68674	-0.25182	-0.55990	0.28114	-0.30160	0.82628	0.75761
303	-1.09862	-1.20890	-1.10585	-0.83980	-0.94628	-0.55990	-0.87528	-0.43957	-0.75204	-0.98944
306	-0.48302	-0.38049	-0.57472	-0.38359	0.05326	-0.55990	-0.23877	-0.43957	0.00555	-0.45591
310	-0.45820	-0.60975	-0.70720	-0.77832	-0.66930	1.17426	-0.74254	0.13991	-0.08915	-0.27807

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
313	-0.43337	-0.46546	-0.54190	-0.67352	-0.22506	-0.55990	-0.26239	-0.38438	-0.43638	-0.39314
314	-1.09862	-1.21110	-1.21524	-0.96556	-1.06136	-0.55990	-0.94674	-0.49476	-0.73626	-0.97898
315	-0.90500	-0.75844	-0.96183	-0.86076	-0.61578	-0.55990	-0.82057	-0.46716	-0.62578	-0.82206
318	-0.80571	-1.03311	-1.06149	-0.88172	-0.89544	-0.55990	-0.80443	-0.49476	-0.45216	-0.83252
321	-0.79082	-0.92544	-0.96304	-0.84539	-0.78973	-0.55990	-0.78978	-0.46716	-0.75204	-0.92667
322	0.67372	0.41788	0.36298	0.85651	2.64512	-0.55990	1.93327	-0.41197	-0.16806	0.02531
323	-0.78585	-0.81484	-0.65189	-0.59947	-0.35485	-0.55990	-0.19632	-0.46716	-0.57843	-0.80114
325	-0.91493	-1.01333	-0.66466	-0.17259	-0.08590	-0.55990	0.57294	-0.43957	-0.75204	-0.98944
328	-0.32415	-0.33069	-0.51881	-0.61554	-0.84058	1.17426	-0.74523	2.43024	0.24230	-0.27807
330	-0.68656	-0.00401	-0.23318	0.46178	0.87618	-0.55990	1.90248	-0.27400	-0.60999	-0.26761
335	-0.66671	-0.82729	-0.85001	-0.73221	-0.65860	-0.55990	-0.62026	-0.49476	-0.42060	-0.74883
338	0.41060	0.81048	0.30282	0.26476	1.21739	0.30718	0.96639	0.85737	1.91532	0.88315
343	-0.43337	-0.24133	-0.56682	-0.66304	-0.23576	1.17426	-0.50187	0.30548	0.54218	-0.04792
344	-0.68160	-0.53065	-0.93934	-0.61274	-0.41239	-0.55990	-0.44895	-0.35678	-0.73626	-0.63375
345	-0.42344	-0.06481	-0.08308	-0.63230	-0.08858	0.74072	-0.59305	0.16751	-0.31011	0.10900
346	-0.41351	-0.44715	-0.61361	-0.44577	0.01713	-0.55990	-0.27644	-0.49476	-0.57843	-0.10022
349	-0.58231	-0.32556	-0.58444	-0.54986	-0.32006	-0.55990	0.05333	-0.32919	-0.75204	-0.57099
352	-0.41351	-0.60975	-0.58748	-0.61134	-0.27992	0.30718	-0.42443	-0.38438	-0.16806	-0.40360
353	-0.49295	-0.41565	-0.60814	-0.79229	-0.40035	0.74072	-0.63939	0.66421	0.46327	-0.06884
356	0.53471	0.27359	-0.00104	-0.48629	-0.03104	-0.12636	-0.29767	0.94015	0.62110	0.69484
363	0.64393	0.41056	0.58358	-0.30604	-0.37760	0.30718	-0.51891	0.60902	1.39448	0.77853
370	0.42053	0.05532	0.00747	0.07333	0.20714	-0.12636	0.33795	-0.32919	0.32122	-0.01653
372	-0.00642	0.01943	-0.27390	0.37584	0.93104	-0.12636	1.11588	-0.49476	-0.31011	-0.38268
373	1.85031	0.87274	0.70087	-0.06710	0.29144	2.04134	-0.20857	0.30548	2.02580	1.60498
379	-0.46812	0.46329	0.04029	1.78641	1.05013	-0.55990	1.58377	-0.49476	-0.72048	-0.58145
382	-0.99933	-1.05948	-1.10281	-0.69029	-0.87670	-0.55990	-0.54372	-0.46716	-0.68891	-0.93713
389	-0.69649	-0.54163	0.09498	0.10686	-0.62782	-0.55990	0.12568	-0.49476	-0.68891	-0.72791
393	-0.60217	-0.44275	-0.44223	0.31506	0.61927	-0.55990	1.38196	-0.49476	-0.40481	-0.20484

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
397	-0.90500	-0.95767	-1.06635	-0.88102	-0.78304	-0.55990	-0.80054	-0.46716	-0.65734	-0.61283
404	-0.89011	-0.94229	-0.92050	-0.77064	-0.66127	-0.12636	-0.56525	-0.46716	-0.72048	-0.84298
406	-0.22983	-0.50721	-0.59234	-0.42550	-0.26922	-0.12636	-0.25611	-0.41197	0.10025	-0.27807
409	0.14251	0.09926	1.62580	0.92778	1.44486	-0.55990	1.29197	-0.46716	0.02134	-0.38268
410	-0.10075	-0.08165	0.15210	-0.27460	-0.47929	-0.55990	-0.10483	-0.43957	-0.72048	-0.19438
415	-0.27451	-0.78408	-0.29335	-0.41642	-0.78170	0.30718	-0.23040	-0.49476	-0.49951	-0.96852
418	0.15244	-0.37097	-0.50848	-0.05033	-0.12203	-0.55990	0.15707	-0.49476	0.03712	-0.20484
419	-0.51777	-0.47791	-0.82266	-0.35774	-0.62782	-0.12636	-0.06537	-0.46716	-0.68891	-0.27807
421	0.04322	-0.19591	0.95550	0.42685	-0.16618	-0.55990	-0.34700	-0.46716	-0.40481	-0.53960
422	-0.90997	-0.85805	-1.01530	-0.52052	0.08404	-0.55990	0.30387	-0.38438	-0.70469	-0.88483
424	-0.78089	-1.01553	-0.95636	-0.76714	-0.92621	-0.55990	-0.82565	-0.46716	-0.62578	-0.69652
427	-1.12841	-1.25431	-1.23773	-0.98582	-1.08946	-0.55990	-0.96886	-0.49476	-0.76783	-1.02083
428	-0.91493	-1.09610	-1.09552	-0.86705	-0.95431	-0.12636	-0.77782	-0.46716	-0.70469	-0.79068
432	-0.95961	-1.15323	-0.88465	-0.66095	-0.43514	-0.55990	-0.61279	-0.49476	-0.76783	-0.87437
433	2.58009	1.19209	1.42222	0.64622	0.31017	-0.12636	0.15139	0.00194	1.55231	1.86652
435	1.02620	1.43453	0.74037	0.82438	1.44486	0.30718	0.83813	1.16090	0.66845	1.86652
438	-0.49295	-0.62147	-0.56985	-0.11181	-0.00561	-0.55990	0.40313	-0.35678	-0.56264	-0.53960
439	0.08294	-0.34753	0.35873	1.42172	0.41722	-0.55990	0.97656	-0.43957	-0.62578	-0.36176
440	0.07797	-0.42884	-0.23197	-0.22080	-0.58233	-0.55990	-0.48393	-0.43957	-0.32590	-0.08976
445	-0.18018	-0.35852	0.16669	-0.64488	-0.52613	-0.55990	-0.53595	-0.43957	-0.34168	-0.39314
447	-1.08373	-1.19059	-1.19701	-0.95298	-1.01319	-0.55990	-0.94465	-0.49476	-0.76783	-0.99990
449	-1.05394	-0.85512	-0.32191	0.26825	-0.78304	-0.55990	-0.27614	-0.49476	-0.76783	-0.90575
466	-1.03408	-1.13858	-1.12105	-0.73920	-0.92755	-0.55990	-0.68245	-0.49476	-0.72048	-0.99990
467	-0.70146	-0.75771	-1.03171	-0.56943	0.46940	-0.55990	0.51733	-0.49476	-0.56264	-0.58145
485	-0.50288	-0.34167	-0.33710	0.89424	0.83470	-0.55990	0.99928	-0.49476	-0.73626	-0.81160
486	-0.99933	-1.15543	-1.15508	-0.94460	-0.99981	-0.12636	-0.88186	-0.49476	-0.73626	-0.89529
490	-0.58727	-0.55409	-0.86702	-0.41083	-0.71480	-0.55990	-0.37271	-0.43957	-0.65734	-0.69652
504	0.08790	-0.10436	-0.06789	-0.10832	0.77047	-0.55990	0.77116	0.41586	-0.18385	-0.60237

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
506	1.14038	0.53141	0.27912	-0.08876	0.39581	-0.55990	0.02552	-0.43957	0.05290	0.06716
509	-0.85536	-0.97232	-0.67256	-0.57711	-0.68402	-0.55990	-0.59186	-0.49476	-0.64156	-0.79068
511	-0.13054	0.91375	0.24387	-0.44646	1.00330	1.60780	-0.14758	1.24369	1.52074	1.72006
514	-0.17025	-0.49549	-0.56682	-0.70426	-0.73888	-0.55990	-0.73088	0.19510	0.44748	-0.30945
515	0.09783	0.32925	-0.07579	-0.48838	0.29278	0.74072	-0.33175	0.96774	0.90520	0.44377
516	0.29145	0.10952	0.00139	-0.14674	-0.08991	-0.55990	-0.14310	-0.49476	-0.57843	0.05670
517	3.24037	1.97947	2.73973	0.51418	0.05728	1.60780	-0.19243	2.18189	4.10919	1.96067
521	0.70847	0.35489	0.13387	-0.33957	-0.29865	0.30718	-0.49469	0.52623	1.89954	0.44377
527	-0.90997	-1.14444	-1.18364	-0.89220	-1.04664	-0.55990	-0.87738	-0.49476	-0.76783	-0.74883
532	-0.10075	-0.69252	-0.61725	-0.75736	-0.64923	-0.55990	-0.70009	-0.46716	-0.73626	-0.96852
533	-0.71139	-0.95034	-1.03475	-0.90757	-0.81515	-0.55990	-0.84329	-0.49476	-0.75204	-0.98944
534	0.83258	0.25235	-0.01927	0.38283	-0.21569	0.30718	-0.24505	-0.21881	0.11603	0.25546
535	-0.79578	-0.94668	-1.00923	-0.75736	-0.80445	-0.12636	-0.46509	-0.46716	-0.65734	-0.68606
536	-0.89507	-0.67787	-0.53886	-0.19495	-0.84459	-0.55990	0.10505	-0.46716	-0.62578	-0.90575
537	-0.37380	-0.85512	-0.93144	-0.82723	-0.85931	-0.55990	-0.69500	-0.43957	-0.42060	-0.55006
538	-0.71635	-0.89175	-0.81415	-0.66724	-0.43648	0.30718	-0.55000	-0.38438	-0.34168	-0.79068
540	0.61414	1.06464	0.17702	0.69163	1.47430	-0.12636	1.49169	0.00194	0.46327	0.78899
542	1.18506	1.20381	0.92511	1.27989	2.73076	0.74072	2.57816	-0.08084	0.14760	0.41238
546	0.12265	-0.45667	-0.31522	-0.35564	-0.65592	-0.55990	-0.22711	-0.46716	-0.68891	-0.75929
548	0.32123	0.21279	0.27972	0.20607	-0.29732	0.30718	0.28623	-0.16362	-0.42060	-0.15253
550	-0.57734	-0.92471	-0.95089	-0.77553	-0.77635	-0.55990	-0.72819	-0.49476	-0.75204	-0.97898
552	-0.86529	-1.02945	-0.64764	-0.11041	-0.42845	-0.55990	-0.10334	-0.49476	-0.70469	-0.98944
553	0.94180	0.71306	1.76922	0.07193	-0.14210	0.30718	-0.30395	0.08472	0.73158	0.75761
556	1.76095	1.09394	0.70694	-0.20683	-0.42979	1.17426	-0.49439	2.01633	2.70448	1.18653
563	-0.23976	-0.22595	-0.34743	0.42615	0.40518	-0.55990	1.00047	-0.16362	-0.42060	0.04624
566	1.31910	1.86741	1.56928	4.81086	2.56216	-0.55990	2.77428	-0.46716	-0.46794	0.26592
568	-0.82557	-0.99429	-0.98917	-0.74059	-0.80177	-0.55990	-0.76347	-0.49476	-0.56264	-0.74883
572	0.19712	-0.31970	0.20862	-0.48699	-0.72684	-0.55990	-0.73596	-0.49476	-0.76783	-0.41407

Br_Num	stnd_br_FreqABF	stnd_br_FreqCA	stnd_br_FreqCAMS	stnd_br_FreqHL	stnd_br_FreqINV	stnd_br_FreqMM	stnd_br_FreqSA	stnd_br_Primary_CG1	stnd_br_Primary_CG2	stnd_br_Primary_SmallBus
574	-0.96458	-1.19718	-1.21706	-0.96765	-1.08277	-0.55990	-0.96139	-0.49476	-0.76783	-0.85344
577	0.28152	0.16299	0.17824	0.42126	-0.03907	-0.55990	0.34363	-0.35678	-0.04180	0.34962
578	1.39357	1.50924	1.00897	-0.16351	0.12016	2.90842	-0.33414	6.04510	2.86232	1.58406
580	-0.89507	-0.93130	-0.58930	-0.54078	-0.75494	-0.55990	-0.71653	-0.49476	-0.68891	-0.86391
581	0.24180	0.18350	-0.01927	-0.36752	-0.42310	-0.12636	-0.46868	0.49864	0.62110	0.62161
582	1.69144	1.78537	1.11411	0.07263	0.74103	1.17426	-0.03607	3.47883	2.81497	2.55697
583	0.38081	-0.42957	-0.32312	-0.02588	-0.62782	-0.55990	-0.39782	-0.46716	-0.16806	-0.63375
591	0.94180	4.15853	2.75127	2.68487	7.57862	-0.12636	6.50158	-0.41197	-0.32590	-0.08976
593	0.90208	0.68450	0.42983	-0.26901	0.47208	1.17426	0.04496	1.38166	2.08894	1.20745
594	0.90705	0.29996	3.21070	1.83951	0.23925	-0.12636	0.70270	-0.41197	-0.05758	-0.48730
595	0.21202	-0.56580	0.66379	-0.37520	-0.80043	-0.55990	-0.22621	-0.41197	-0.56264	-0.86391
596	-0.47309	-0.84633	0.66197	-0.66025	-0.79374	-0.55990	-0.77184	-0.49476	-0.70469	-0.97898
602	-1.08869	-1.18912	-1.18121	-0.95438	-1.05333	-0.12636	-0.93837	-0.38438	-0.73626	-0.95806
603	0.34606	0.25967	0.94759	0.11385	-0.11935	0.30718	-0.01095	-0.38438	0.17917	0.22408
618	2.00421	1.77512	1.83060	0.14040	0.39313	2.04134	-0.28391	1.63001	2.62557	2.53604
620	0.39570	0.46256	0.30646	0.15437	0.09340	-0.55990	0.03868	-0.43957	-0.24698	0.42284
623	-0.24472	-0.18053	-0.49632	-0.46393	0.65540	-0.55990	-0.03128	-0.43957	0.08447	0.36008
627	0.40563	0.18569	-0.21556	-0.49048	0.02516	1.17426	-0.43998	0.38826	1.22086	0.38100
633	2.14818	1.07563	1.02781	0.60500	0.22587	-0.12636	-0.08271	-0.19122	0.36857	0.67392
634	-0.41351	-0.41712	-0.23136	-0.65746	-0.46859	-0.55990	-0.64956	-0.19122	0.06868	-0.07930
638	-0.58231	-0.63685	-0.61179	-0.06919	1.55191	0.30718	0.78163	-0.41197	-0.29433	-0.79068
640	0.04322	0.97674	0.75860	0.40728	0.81597	-0.55990	0.37024	-0.24641	-0.23120	0.44377
642	0.09287	0.27945	0.46204	0.19280	-0.08055	-0.55990	-0.20349	-0.43957	-0.48373	-0.21530
645	0.98648	0.28604	0.18310	0.47086	-0.08991	-0.55990	0.10834	-0.38438	-0.29433	0.12993
651	1.89499	2.43140	3.42218	3.37723	2.22363	0.30718	2.62390	-0.19122	0.00555	0.80992
657	-0.61210	-0.24499	-0.70720	-0.49118	0.77984	-0.55990	0.34483	-0.32919	-0.10493	-0.56053
659	0.76308	0.52336	0.29674	1.81366	0.68751	-0.55990	2.48966	-0.49476	-0.35746	-0.25714
667	-0.05607	-0.69618	1.12505	-0.12299	-0.54887	-0.55990	-0.68065	-0.49476	-0.68891	-0.60237

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
668	-0.49295	-0.57752	-0.32738	1.25055	2.93682	-0.55990	1.52667	-0.49476	-0.64156	-0.52914
669	1.54251	1.19135	0.64617	1.97645	1.82889	0.74072	3.62576	-0.43957	0.35278	1.15514
670	-0.81564	-0.79506	-0.89194	-0.78461	-0.87403	-0.55990	-0.78589	-0.41197	-0.62578	-0.76975
671	-0.80571	-0.56287	-0.76067	-0.53100	-0.48197	-0.55990	-0.30634	-0.49476	-0.70469	-0.95806
672	0.54464	1.10932	1.48785	3.51696	2.22095	-0.55990	3.29390	-0.38438	-0.43638	0.11946
676	-0.40855	0.04946	-0.20037	-0.54218	-0.34415	0.74072	-0.63132	1.65760	0.40013	0.53792
689	-1.08869	-1.18693	-1.12226	-0.95927	-1.00115	-0.55990	-0.94195	-0.46716	-0.72048	-0.98944
690	2.52052	1.78464	0.46325	1.17510	0.45335	-0.55990	1.42502	-0.49476	0.03712	0.45423
692	-0.86529	-1.00967	-0.07214	-0.08526	-0.37359	-0.55990	-0.12157	-0.43957	-0.75204	-0.94760
713	0.13258	0.19668	0.44380	-0.00352	-0.08055	-0.12636	-0.43669	-0.49476	0.77893	0.14039
717	-0.81068	-0.72328	-0.75946	-0.75527	-0.75628	-0.55990	-0.75061	-0.43957	-0.70469	-0.70698
720	-0.69153	-0.55628	-0.13413	0.25288	-0.33077	-0.55990	0.32061	-0.49476	-0.75204	-0.90575
729	-1.03408	-1.19938	-1.18182	-0.91595	-1.04129	-0.55990	-0.92013	-0.49476	-0.76783	-0.97898
733	-0.48302	-0.40247	-0.40030	-0.02937	-0.42979	-0.55990	-0.15566	-0.49476	-0.64156	-0.58145
735	-0.84046	-0.78041	-0.84758	-0.81814	-0.83924	-0.12636	-0.83193	-0.13603	-0.40481	-0.58145
736	-0.24969	0.09780	0.48877	0.52535	0.03185	-0.55990	0.09279	-0.49476	-0.45216	-0.44545
739	-1.05394	-1.20231	-1.19762	-0.95717	-1.06404	-0.55990	-0.95541	-0.49476	-0.70469	-0.97898
743	-1.05890	-1.16495	-1.17452	-0.92574	-1.03995	-0.55990	-0.92162	-0.49476	-0.70469	-0.97898
747	-0.63195	-0.60096	-0.61422	-0.43389	-0.53549	-0.55990	-0.15984	-0.49476	-0.65734	-0.44545
752	-0.43337	-0.33069	0.65529	-0.37171	-0.67599	-0.55990	0.18547	-0.43957	-0.70469	-0.97898
757	-1.02912	-1.12686	-1.15508	-0.87054	-0.93424	-0.55990	-0.82326	-0.46716	-0.68891	-0.93713
758	0.04322	-0.14611	-0.14932	-0.44367	-0.50204	-0.55990	-0.59186	0.05713	0.22652	0.00439
764	0.09783	0.21719	1.56442	0.90961	0.39581	-0.55990	0.62018	-0.43957	-0.15228	-0.15253
768	-0.20004	-0.56214	-0.68896	-0.21521	-0.58634	-0.12636	-0.31022	-0.46716	-0.70469	0.12993
772	-0.46316	-0.61927	-0.57532	-0.69518	-0.67332	-0.55990	-0.79665	-0.32919	-0.59421	-0.79068
783	-0.35891	-0.62733	-0.51091	-0.64278	-0.64655	-0.55990	-0.76287	-0.41197	0.10025	-0.52914
784	-0.22486	-0.71962	-0.56074	-0.51493	-0.03505	-0.12636	-0.14400	-0.41197	-0.43638	-0.79068
795	-1.02912	-1.18326	-1.19397	-0.93272	-1.01720	-0.55990	-0.83761	-0.49476	-0.76783	-0.85344

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
796	-0.69649	-0.95474	-0.88525	-0.43249	-0.24781	-0.55990	-0.21874	-0.46716	-0.72048	-0.99990
800	0.00847	0.39444	0.04454	0.13761	1.39535	-0.55990	0.77923	-0.21881	-0.35746	0.46469
802	-1.02912	-1.12540	-1.16298	-0.86705	-1.02925	-0.55990	-0.86990	-0.49476	-0.75204	-0.76975
803	1.34889	0.90130	0.34475	-0.29067	-0.05245	1.17426	-0.30933	1.49203	1.23664	1.77236
807	1.75102	1.76340	1.33896	0.63854	1.83558	2.90842	1.43189	1.60241	2.75183	1.74098
808	-0.61210	-0.74013	-0.71449	-0.42201	-0.69606	-0.55990	-0.62116	-0.46716	-0.57843	-0.45591
822	0.31627	0.06704	-0.10556	-0.25084	-0.23443	0.30718	-0.44446	-0.35678	0.70001	0.32869
825	0.84747	-0.00914	-0.05391	-0.42900	-0.56359	0.30718	-0.51502	-0.08084	0.66845	0.20316
826	0.65386	0.42447	1.04179	2.41520	1.35120	-0.55990	2.64841	-0.35678	-0.46794	1.38529
834	1.37371	1.59054	1.64889	0.28712	0.29010	0.30718	-0.07493	0.47104	2.08894	2.26405
835	-0.62202	-0.84926	-0.70112	-0.75457	-0.84058	-0.55990	-0.77662	-0.38438	-0.70469	-0.89529
836	0.49003	2.75661	0.47662	1.61594	2.15405	-0.12636	2.63526	-0.02565	-0.04180	0.66346
839	0.10776	-0.29260	-0.25810	-0.35424	-0.19295	-0.55990	-0.35447	-0.41197	-0.05758	0.03577
840	-0.11068	1.20747	0.40977	2.28245	2.77893	-0.55990	2.69984	-0.43957	-0.54686	-0.07930
843	1.47300	0.63762	0.58905	0.73635	0.06263	1.17426	0.25573	-0.46716	0.21073	1.46898
844	-0.23479	-0.66102	-0.03811	-0.34027	-0.54887	-0.12636	-0.41158	-0.43957	-0.56264	-0.49776
849	-0.55252	-0.76650	-0.87614	-0.68261	-0.77501	-0.55990	-0.73417	-0.46716	-0.46794	-0.37222
852	-0.17025	-0.24426	0.05244	1.13178	0.00777	-0.55990	1.23547	-0.46716	-0.40481	-0.07930
853	-0.91493	-1.08438	-1.04204	-0.82862	-1.01586	-0.55990	-0.88186	-0.49476	-0.65734	-0.91621
854	0.80279	1.62057	1.61790	2.44663	2.28250	-0.55990	2.47322	-0.46716	-0.04180	0.59023
855	-0.99933	-1.05801	-0.66283	-0.47581	-0.07386	-0.55990	0.99001	-0.49476	-0.75204	-0.97898
862	-0.68160	-0.89907	-0.90835	-0.78251	-0.76297	-0.55990	-0.80024	-0.49476	-0.59421	-0.70698
864	0.75315	0.03554	-0.14689	0.08521	-0.33077	-0.12636	-0.37361	-0.38438	0.13182	0.16131
876	0.09783	-0.14391	-0.22224	-0.07967	-0.12470	-0.55990	0.49969	-0.21881	-0.38903	0.09854
877	-1.07380	-0.90859	-1.15386	-0.70706	-0.64522	-0.55990	-0.22023	-0.49476	-0.76783	-0.99990
879	0.78790	0.62810	0.63462	1.03327	3.35698	0.30718	2.04837	-0.35678	0.06868	-0.06884
883	-0.03621	-0.13805	0.23293	-0.30254	-0.46324	-0.12636	-0.54462	-0.19122	-0.21541	0.25546
884	-0.50288	-0.85146	-0.50665	-0.76435	-0.87938	-0.55990	-0.81071	-0.46716	-0.76783	-0.93713

Br_Num	stnd_br_ FreqABF	stnd_br_ FreqCA	stnd_br_ FreqCAMS	stnd_br_ FreqHL	stnd_br_ FreqINV	stnd_br_ FreqMM	stnd_br_ FreqSA	stnd_br_ Primary_CG1	stnd_br_ Primary_CG2	stnd_br_Primary _SmallBus
887	0.35599	0.39591	-0.30307	0.21026	0.04122	0.30718	0.15886	-0.08084	-0.56264	1.17606
891	-0.88018	-0.64637	-0.24655	0.19350	0.24728	-0.55990	0.54185	-0.49476	-0.65734	-0.72791
895	0.28152	0.14761	0.00018	-0.32490	-0.31337	-0.55990	-0.34461	-0.41197	0.21073	0.38100
898	0.55457	0.36222	0.04272	-0.26132	0.24594	1.17426	-0.28690	0.05713	0.84206	0.68438
899	-0.79082	-1.00601	-1.01044	-0.92364	-0.92621	-0.55990	-0.88485	-0.32919	-0.51529	-0.79068
919	1.10066	2.03441	1.20101	0.00626	1.54923	4.64258	0.23361	2.65100	2.67292	3.16373
925	1.98435	2.22997	1.33349	0.37934	1.60945	3.77550	0.18547	2.18189	1.72592	3.38342
926	-0.49791	-0.21642	0.14056	-0.03077	-0.38697	-0.55990	-0.34640	-0.49476	-0.51529	-0.38268
928	0.38577	0.08242	-0.08369	-0.39267	-0.23844	0.30718	-0.46539	0.38826	0.66845	0.48561
930	0.71343	0.53361	0.34596	0.20188	0.32623	0.30718	-0.01514	-0.49476	0.52640	0.71576
932	-0.42841	-0.57533	-0.91139	-0.56873	-0.69071	-0.55990	-0.44656	-0.46716	-0.72048	-0.43499
934	1.37868	1.13495	2.42676	2.04491	1.49571	-0.12636	1.57420	-0.13603	0.62110	1.06099
936	1.77584	1.06098	0.73550	1.38259	0.64603	0.74072	0.53916	-0.30160	0.51062	1.17606
937	-0.59224	-0.84780	-0.79531	-0.74059	-0.83389	-0.55990	-0.79127	-0.41197	-0.45216	-0.66514
939	-0.59720	-0.83901	-0.63002	-0.51563	-0.76029	-0.55990	-0.69261	-0.46716	-0.64156	-0.66514
958	-0.46316	-0.30945	-0.57289	-0.75527	-0.59838	1.60780	-0.73417	2.18189	0.14760	0.48561
971	-0.56245	-0.43616	-0.55345	-0.67073	-0.57162	0.30718	-0.72998	1.98873	0.19495	-0.37222
973	-0.37380	-0.20983	-0.44588	-0.67632	-0.54352	2.47488	-0.69889	1.85076	0.54218	0.17177
975	1.31414	1.93406	1.22714	-0.00492	1.04879	1.17426	0.08591	4.55500	3.25690	3.17419
976	-0.49791	-0.28747	-0.62212	-0.80347	-0.40704	-0.12636	-0.74912	0.88496	-0.16806	-0.18392
977	-0.41351	0.04140	-0.59112	-0.80697	-0.28394	0.30718	-0.74673	1.18850	-0.23120	0.48561
980	-0.39862	-0.80385	-0.90956	-0.76225	-0.89544	-0.12636	-0.78768	-0.49476	-0.65734	-0.11069
981	0.52974	0.13296	0.07918	0.12573	-0.10329	0.74072	-0.15207	-0.32919	-0.04180	0.23454
985	3.45881	2.22778	1.80993	0.40030	0.70491	2.47488	0.17680	1.60241	4.15654	2.74527
989	0.27159	-0.14318	-0.02170	0.00067	-0.38295	-0.55990	-0.08928	-0.32919	-0.01023	0.27639

Appendix B – Data envelopment analysis input

Br_Num	Br_Cstmr_ Satisfaction	Total_Txn _Volume	Total_Sales _Volume	Total _NII	Total_ Human_ Cost	Total_ Rental_Cost	Total_ Operational_ Cost
15	0.963	1.256	1.59	1.949	1.496	0.512	1.041
20	1.02	0.72	0.131	0.048	0.599	0.806	0.678
45	1.017	0.993	0.57	0.23	1.392	1.805	1.225
49	0.975	0.456	0.379	0.278	0.939	0.937	1.56
62	1.057	0.703	0.312	0.093	0.606	0.494	0.656
63	1.095	0.764	1.23	0.11	0.541	0.267	0.599
79	1.039	1.137	1.106	0.528	1.448	0.991	1.336
83	1.002	1.015	1.681	4.586	1.067	7.042	1.616
84	1.026	0.859	0.717	0.55	1.428	1.867	1.122
88	0.966	1.095	1.332	4.061	1.464	1.617	1.242
90	1.017	0.846	0.864	1.277	0.868	0.918	0.815
92	1.025	0.847	0.219	0.908	0.713	0.858	0.711
93	1.071	1.009	1.312	3.413	0.91	0.86	0.974
124	1.061	0.989	0.869	0.783	0.946	1.52	0.759
148	1.065	0.438	1.338	0.261	1.06	0.478	1.086
185	0.959	0.763	0.193	0.309	0.566	1.291	0.665
203	1.018	2.164	0.335	0.125	1.697	5.243	2.16
214	1.013	1.018	0.485	0.649	0.644	0.638	0.716
216	1.018	1.419	2.39	4.028	1.885	1.994	1.379
221	0.947	0.495	0.412	0.259	1.067	0.323	0.887
229	0.994	0.505	0.69	1.143	0.595	0.246	0.53
238	0.954	1.203	0.359	0.213	0.842	1.307	0.932
243	1.035	0.414	0.348	0.179	1.067	0.356	0.86
251	0.965	0.427	0.481	0.239	1.067	0.855	1.155
289	1.022	0.736	0.846	0.278	0.562	0.627	0.815
303	0.926	1.508	1.311	2.878	0.971	1.234	0.816
313	0.984	1.491	1.613	2.437	0.99	0.642	0.946
314	0.911	1.442	1.119	2.458	1.149	1.867	1.152
315	1.089	0.423	0.478	0.193	0.567	0.333	0.462
318	1.033	0.936	0.198	0.118	0.559	0.345	0.785
321	0.93	0.334	0.255	0.142	1.067	0.742	0.901
323	1.067	1.443	1.74	1.84	1.132	0.503	1.119
335	1.033	1.069	1.551	1.925	0.928	0.561	0.812
344	0.949	0.737	1.048	0.69	0.557	0.394	0.524
346	0.963	1.53	1.391	1.107	0.945	0.59	1.028
349	1.021	1.657	1.834	1.164	1.313	0.582	1.032
352	0.961	1.242	1.647	4.562	1.39	0.685	1.064
382	0.963	1.546	0.917	1.372	1.025	3.851	1.036
397	1.01	0.895	1.393	1.914	1.264	2.009	1.348
404	1.018	1.115	0.919	0.47	0.681	0.323	0.706
406	1.06	1.495	1.525	0.49	0.74	1.033	0.969
415	1.022	0.987	1.146	0.199	0.774	0.392	1.153
419	0.912	1.797	1.218	1.519	1.077	0.949	1.292
422	0.915	0.775	1.584	0.272	0.604	0.209	0.583
427	0.926	2.032	1.436	1.709	1.31	0.7	1.138

Br_Num	Br_Cstmr_ Satisfaction	Total_Txn _Volume	Total_Sales _Volume	Total _NII	Total_ Human_ Cost	Total_ Rental_Cost	Total_ Operational_ Cost
428	0.885	1.303	1.75	1.083	1.28	0.998	1.591
432	0.866	1.03	2.187	1.056	0.871	0.772	1.17
440	1.033	0.506	0.262	0.03	1.165	1.818	1.024
445	0.843	0.846	0.837	0.973	0.893	0.4	0.791
449	1.033	0.589	0.641	0.188	0.764	0.477	0.627
466	1.047	0.898	1.423	0.483	0.688	0.311	0.825
486	1.01	2.222	3.125	1.16	1.151	1.436	1.525
490	1.036	0.581	0.648	0.12	0.602	0.21	0.73
509	0.983	1.571	2.971	0.694	1.461	0.807	1.215
514	0.979	1.548	0.982	1.647	1.211	1.494	1.138
527	1.055	1.034	0.666	0.149	1.159	0.923	0.842
532	1.009	0.465	0.761	0.227	1.067	0.784	1.09
533	1.029	0.237	0.299	0.095	1.067	0.224	0.802
535	0.924	0.555	1.242	0.523	0.772	0.236	1.042
536	0.97	0.52	0.598	0.361	0.932	1.742	1.257
537	0.991	1.401	2.238	1.055	1.232	1.967	1.309
538	0.973	0.686	1.106	0.654	0.513	0.301	0.962
546	1.013	0.89	0.791	0.38	1.63	1.485	1.16
550	1.088	0.293	0.34	0.112	1.067	0.571	0.819
552	0.898	0.756	0.669	0.247	0.532	0.677	0.415
572	1.03	0.923	0.591	0.269	1.005	0.621	0.862
574	1.033	0.745	1.142	0.437	0.654	0.932	0.802
580	0.956	1.117	2.332	0.996	1.368	0.423	1.078
583	1.011	1.411	0.611	0.663	0.723	0.357	0.756
596	0.952	0.416	0.24	0.108	1.067	0.61	0.792
602	0.983	1.458	1.314	5.392	1.927	1.581	2.58
670	1.041	0.343	0.292	0.161	1.067	0.569	1.035
671	1.049	1.398	1.965	0.831	0.928	0.334	1.285
689	0.983	1.243	0.729	0.402	1.067	1.131	1.637
692	1.018	0.617	0.678	0.225	0.577	0.211	0.581
717	1.06	0.335	0.68	0.364	0.591	0.321	0.475
729	1.026	1.476	1.634	4.758	1.4	1.004	1.512
735	1.017	1.194	1.032	0.535	1.417	1.761	1.058
739	1.052	1.114	0.68	0.778	1.029	1.664	1.103
743	0.949	1.359	0.856	0.37	0.624	0.404	0.802
747	0.961	1.411	1.224	1.882	1.312	0.793	1.029
757	1.009	1.613	1.672	0.558	0.989	2.383	1.153
768	1.104	0.49	0.058	0.047	0.582	0.201	0.556
772	1.01	0.587	0.516	0.131	0.568	0.273	0.82
783	0.965	1.301	1.194	2.164	1.219	1.271	1.048
784	0.978	0.893	0.746	0.287	0.576	0.21	0.562
795	0.961	1.03	0.219	0.026	0.519	0.133	0.675
796	0.991	0.77	0.874	0.289	0.585	0.285	0.61
808	1.088	1.145	1.557	0.796	1.576	0.513	1.699
835	1.048	0.479	0.276	0.184	0.586	1.297	0.524
849	1.047	1.153	1.197	0.538	1.006	1.023	1.053
853	1.063	1.644	1.352	1.097	1.451	1.3	1.068
862	0.926	0.868	1.106	0.797	1.289	0.821	0.903
877	0.907	0.707	1.308	0.236	0.541	0.638	0.737
884	1.059	0.604	0.351	0.097	1.067	0.677	0.898
899	1.038	1.232	1.048	1.228	0.855	2.081	1.021

Br_Num	Br_Cstmr_ Satisfaction	Total_Txn _Volume	Total_Sales _Volume	Total _NII	Total_ Human_ Cost	Total_ Rental_Cost	Total_ Operational_ Cost
932	1.029	0.643	0.233	0.725	0.523	0.243	0.341
937	1.017	0.847	0.343	1.166	2.006	3.33	1.274
939	0.982	1.192	1.01	2.53	1.37	0.53	1.265
980	0.974	1.585	0.894	3.743	1.369	0.646	1.685

Appendix C – Decision tree induction input

Br_Num	Total_ HR_Cost	Total_ Rent	Total_ Operational_Cost	Cnt_ Banks	Avg_ Income	Litracy _Pop	UnEmployment _Pop	Avg_HSE _Price	Sub_ Group	Efficiency
15	1.496	0.512	1.041	6	63,432	887	1183	467,342	E	0
20	0.599	0.806	0.678	3	100,604	399	423	277,445	C	0
45	1.392	1.805	1.225	19	193,749	259	430	722,077	B	0
49	0.939	0.937	1.56	3	222,049	70	77	857,125	S	0
62	0.606	0.494	0.656	19	265,837	66	73	1,079,963	S	0
63	0.541	0.267	0.599	6	29,265	676	822	69,452	C	1
79	1.448	0.991	1.336	12	51,063	74	92	715,750	H	0
83	1.067	7.042	1.616	34	295,541	89	144	716,374	S	1
84	1.428	1.867	1.122	20	150,892	546	556	841,030	G	0
88	1.464	1.617	1.242	6	139,352	702	1230	624,308	B	0
90	0.868	0.918	0.815	3	122,849	905	1499	789,196	B	0
92	0.713	0.858	0.711	36	177,319	20	50	1,350,562	S	0
93	0.91	0.86	0.974	16	127,421	115	250	598,694	C	1
124	0.946	1.52	0.759	6	82,040	144	306	274,424	D	0
148	1.06	0.478	1.086	8	78,612	592	1023	479,216	C	0
185	0.566	1.291	0.665	34	-	1	5	-	S	0
203	1.697	5.243	2.16	30	77,563	483	1787	270,346	H	0
214	0.644	0.638	0.716	23	76,584	92	263	323,042	C	0
216	1.885	1.994	1.379	26	185,863	24	47	778,918	C	0
221	1.067	0.323	0.887	3	195,885	3,271	3813	344,235	E	0
229	0.595	0.246	0.53	23	72,309	415	817	254,517	H	1
238	0.842	1.307	0.932	21	60,526	1,823	3538	278,437	H	0
243	1.067	0.356	0.86	3	69,743	2,115	2288	280,269	E	0
251	1.067	0.855	1.155	3	77,860	2,480	1945	246,777	E	0
289	0.562	0.627	0.815	49	102,994	663	1141	596,635	C	0
303	0.971	1.234	0.816	6	57,340	141	200	269,702	D	1
313	0.99	0.642	0.946	49	87,419	62	112	616,417	C	1

Br_Num	Total_ HR_Cost	Total_ Rent	Total_ Operational_Cost	Cnt_ Banks	Avg_ Income	Litracy _Pop	UnEmployment _Pop	Avg_HSE _Price	Sub_ Group	Efficiency
314	1.149	1.867	1.152	48	102,994	663	1141	596,635	C	0
315	0.567	0.333	0.462	2	62,037	1,116	2107	428,158	E	0
318	0.559	0.345	0.785	16	157,530	416	591	353,004	C	0
321	1.067	0.742	0.901	3	70,051	1,163	1307	384,290	E	0
323	1.132	0.503	1.119	3	81,329	270	405	586,028	F	0
335	0.928	0.561	0.812	10	110,773	1,680	2351	409,108	C	0
344	0.557	0.394	0.524	3	52,507	2,431	2081	219,036	C	1
346	0.945	0.59	1.028	5	88,991	1,001	1338	392,622	F	0
349	1.313	0.582	1.032	3	54,249	452	828	353,631	D	0
352	1.39	0.685	1.064	10	165,921	1,274	1475	1,458,736	A	1
382	1.025	3.851	1.036	12	133,347	2,367	5649	765,282	E	0
397	1.264	2.009	1.348	11	159,736	400	654	604,990	B	0
404	0.681	0.323	0.706	21	60,526	1,823	3538	278,437	H	0
406	0.74	1.033	0.969	16	157,530	416	591	353,004	C	1
415	0.774	0.392	1.153	4	63,395	142	137	247,873	B	0
419	1.077	0.949	1.292	6	88,449	364	368	513,513	D	0
422	0.604	0.209	0.583	4	21,349	1,546	1330	94,970	E	1
427	1.31	0.7	1.138	4	51,037	723	1121	414,380	E	1
428	1.28	0.998	1.591	6	88,449	364	368	513,513	D	0
432	0.871	0.772	1.17	5	88,991	1,001	1338	392,622	F	1
440	1.165	1.818	1.024	4	128,926	457	587	420,988	B	0
445	0.893	0.4	0.791	3	67,960	919	444	343,963	S	0
449	0.764	0.477	0.627	12	51,063	74	92	715,750	H	0
466	0.688	0.311	0.825	8	69,536	2,292	4004	261,800	F	0
486	1.151	1.436	1.525	48	90,140	1,732	5228	202,100	H	1
490	0.602	0.21	0.73	0	126,531	410	512	185,107	S	0
509	1.461	0.807	1.215	5	62,158	393	653	315,779	F	0
514	1.211	1.494	1.138	19	244,432	950	978	698,618	B	0
527	1.159	0.923	0.842	75	61,398	73	178	131,426	H	0
532	1.067	0.784	1.09	3	49,511	1,422	1341	96,981	E	0
533	1.067	0.224	0.802	3	47,897	316	256	281,261	E	0
535	0.772	0.236	1.042	12	38,811	973	2048	245,959	J	0
536	0.932	1.742	1.257	11	376,721	3,148	1692	845,033	S	0

Br_Num	Total_ HR_Cost	Total_ Rent	Total_ Operational_Cost	Cnt_ Banks	Avg_ Income	Litracy _Pop	UnEmployment _Pop	Avg_HSE _Price	Sub_ Group	Efficiency
537	1.232	1.967	1.309	16	157,530	416	591	353,004	C	0
538	0.513	0.301	0.962	7	75,498	177	298	1,005,252	S	1
546	1.63	1.485	1.16	3	57,198	1,037	1674	403,452	E	0
550	1.067	0.571	0.819	3	57,735	564	659	230,887	E	0
552	0.532	0.677	0.415	3	49,525	760	984	246,339	E	1
572	1.005	0.621	0.862	4	60,861	315	405	437,871	F	0
574	0.654	0.932	0.802	5	83,324	73	104	589,275	S	0
580	1.368	0.423	1.078	5	76,792	531	950	503,097	F	0
583	0.723	0.357	0.756	11	91,483	449	796	343,795	E	1
596	1.067	0.61	0.792	3	77,927	1,220	1054	377,240	E	0
602	1.927	1.581	2.58	43	101,698	42	311	978,935	B	0
670	1.067	0.569	1.035	3	134,942	287	270	1,614,233	A	0
671	0.928	0.334	1.285	5	285,927	282	306	395,816	E	1
689	1.067	1.131	1.637	8	236,581	134	403	583,484	B	0
692	0.577	0.211	0.581	4	49,561	1,269	1534	158,935	F	0
717	0.591	0.321	0.475	1	60,647	229	408	103,368	F	0
729	1.4	1.004	1.512	13	87,228	1,004	2396	198,027	E	0
735	1.417	1.761	1.058	11	114,830	417	4106	570,722	H	0
739	1.029	1.664	1.103	15	23,084	4	2	-	S	0
743	0.624	0.404	0.802	48	111,274	1,187	3042	370,065	D	1
747	1.312	0.793	1.029	5	70,651	97	104	802,577	B	0
757	0.989	2.383	1.153	46	124,442	290	432	979,571	S	0
768	0.582	0.201	0.556	16	86,381	262	579	1,335,576	S	1
772	0.568	0.273	0.82	3	73,733	836	2295	397,464	B	0
783	1.219	1.271	1.048	13	110,452	1,028	1810	388,537	C	0
784	0.576	0.21	0.562	10	142,198	219	475	292,409	E	1
795	0.519	0.133	0.675	48	75,853	316	775	723,074	S	1
796	0.585	0.285	0.61	5	173,213	40	73	166,666	E	0
808	1.576	0.513	1.699	12	87,228	1,004	2396	198,027	E	0
835	0.586	1.297	0.524	3	130,830	1,898	1718	215,506	E	0
849	1.006	1.023	1.053	11	32,832	178	315	308,347	E	0
853	1.451	1.3	1.068	16	118,300	1,500	2875	363,431	C	0
862	1.289	0.821	0.903	16	77,155	841	1964	285,074	C	0

Br_Num	Total_ HR_Cost	Total_ Rent	Total_ Operational_Cost	Cnt_ Banks	Avg_ Income	Litracy _Pop	UnEmployment _Pop	Avg_HSE _Price	Sub_ Group	Efficiency
877	0.541	0.638	0.737	1	43,801	1,801	1122	104,541	E	0
884	1.067	0.677	0.898	3	56,559	298	566	357,233	E	0
899	0.855	2.081	1.021	12	105,462	-	1	417,268	S	0
932	0.523	0.243	0.341	67	309,910	515	1471	286,421	H	1
937	2.006	3.33	1.274	14	79,254	2	7	-	S	0
939	1.37	0.53	1.265	4	99,508	1,480	2972	274,669	H	0
980	1.369	0.646	1.685	59	58,638	32	91	616,333	S	1

References

AHVENLAMPI, T. & KORTELA, U. 2005. Clustering algorithms in process monitoring and control application to continuous digesters. *Informatica*, 29(1):99–109, May.

ALI, A.I. & SEIFORD, L.M. 1993. The mathematical programming approach to efficiency analysis. (*In* Fried, H.O., Knox Lovell, C.A. & Schmidt, S.S., *eds.* The measurement of productive efficiency; techniques and applications. Oxford: Oxford University Press. p. 120-159.)

BAKER, K.R. 2006. Optimization modeling with spreadsheets. Belmont, CA: Thomson Brooks/Cole.

BANKER, R., CHARNES, A., & COOPER, W.W. 1984. Some models for estimating technical and scale inefficiencies in data envelopment analysis. *Management science*, 30(9):1078–1092, Sep.

BERKHIN, P. 2002. Survey of clustering data mining techniques. http://www.ee.ucr.edu/~barth/EE242/clustering_survey.pdf Date of access: 20 Oct. 2007.

BERRY, M.J.A. & LINOFF, G.S. 2000. Mastering data mining: the art and science of customer relationship management. 2nd ed. New York: Wiley Computer.

BERRY, M.J.A. & LINOFF, G.S. 2004. Data mining techniques: for marketing, sales and customer relationship management. 2nd ed. Indianapolis, IN: Wiley.

BREWERTON, P. & MILLWARD, L. 2001. Organizational research methods. London: Sage.

CABENA, P., HADJINIAN, P., STADLER, R., VERHEES, J. & ZANASI, A. 1998. Discovering data mining from concept to implementation. Upper Saddle River, NJ: Prentice-Hall.

CAMERON, F., CORNISH, C. & NELSON, W. 2006. A new methodology for segmenting consumers for financial services. *Journal of financial services marketing,* 10(3):260-271, Feb.

CAVELL, D.J. 2002. Branch profitability: strategies for making branches pay. London: Lafferty.

CAVELL, D.J., KANE, M.L. & RYAN, T.J. 2002. Branch network development in a multi-channel world. London: Lafferty.

CHARNES, A., COOPER, W.W., LEWIN, A.Y. & SEIFORD, L.M. 1994. Data envelopment analysis theory, methodology and applications. London: Kluwer Academic.

CHARNES, A., COOPER, W.W., RHODES, E.L. 1978. Measuring the efficiency of decision making units. *European journal of operational research*, 2(6):429–444, Nov.

CHEN, Y-L., HSU, C.L., & CHOU, S.C. 2003. Constructing a multi-valued and multi-labeled decision tree. *Expert systems with applications,* 25(2), 199–209, Aug.

CHEN. Y-L ., HU. H-W. & TANG. K. 2009. Constructing a decision tree from data with hierarchical class labels. *Expert systems with applications,* 36(3,1):4838-4847, Apr.

CHOPOORIAN, J.A., WITHERELL, R., KHALIL, O.E.M. & AHMED, M. 2001. Mind your business by mining your data. *SAM advanced management journal*, 66(2):45, Mar.

COOK, W.D. & SEIFORD, L.M. 2009, Data envelopment analysis (DEA) - thirty years on. *European journal of operational research*, 129(1):1-17, Jan.

COOK, W.D. & ZHU, J. 2007. Classifying inputs and outputs in data envelopment analysis. *European journal of operational research*, 180(2):692–699, Jul.

COOPER, W.W., SEIFORD, L.M. & TONE, K. 2006. Introduction to data envelopment analysis and its uses. New York, NY: Springer.

COOPER, W.W., SEIFORD, L.M. & TONE, K. 2007. Data envelopment analysis: a comprehensive text with models, applications, references and DEA-solver software. 2nd ed. New York, NY: Springer.

DASZYKOWSKI, M., WALCZAK1, B. & MASSART, D.L. 2002. Representative subset selection. *Analytica chimica acta*, 468(1):91–103, Sep.

DATA MINERS, INC. 2004. Data mining techniques: theory and practice course notes. Johannesburg: SAS Institute Inc. (BDMT/59933).

DAWSON, C. 2007. A practical guide to research methods. 3rd ed. Oxford: Howtobooks.

DE LEONE, R. & LAZZARI, C. 1998. Measuring efficiency using data envelopment analysis. Camerino, Italy: Università degli Studi di Camerino Dipartimento di Matematica e Fisica; Operations Research Group. (Technical Report no. OR-CAM-1998-03) (Unpublished)

DÍEZ, J., DEL COZ, J.J., BAHAMONDE, A., SAÑUDO, C., OLLETA, J.L., MACIE, S., CAMPO, M.M., PANEA, B. & ALBERTÍ, P. 2006. Identifying market segments in beef: breed, slaughter weight and ageing time implications. *Meat science*, 74(4):667-675, Dec.

D' SOUZA, R., KRASNODEBSKI, M. & ABRAHAMS, A. 2007. Using decision tree induction to discover profitable locations to sell pet insurance for a startup company. *Journal of database marketing & customer strategy management*, 14(4):281–288, Jul.

DUNHAM, M.H. 2003. Data mining introductory and advanced topics. Upper Saddle River, NJ: Prentice Hall.

EASTERBY-SMITH, M., THORPE, R. & LOWE, A. 2002. Management research. 2nd ed. London: Sage.

EVANS, B. & FISHER, D. 1994. Overcoming process delays with decision tree induction. *IEEE expert*, 9(1):60-66, Feb.

FARRELL, M.J. 1957. The measurement of productive efficiency. *Journal of the royal statistical society*, 120(3):253–281.

FATTI, L.P. & CLARKE, K. 1998. Use of data envelopment analysis and regression for establishing manpower requirements in a bank, *Orion*, 14(1/2):57-66.

FERNANDEZ, G. 2003. Data mining using sas applications. Florida: Chapman & Hall/CRC

FLOREZ-LOPEZ, R. 2007. Modelling of insurers' rating determinants: an application of machine learning techniques and statistical models. *European journal of operational research,* 183(3):1488–1512, Dec.

FOX, W. & BAYAT, M.S. 2007. A guide to managing research. Cape Town: Juta.

GELBARD, R., GOLDMAN, O. & SPIEGLER, I. 2007. Investigating diversity of clustering methods: an empirical comparison. *Data & knowledge engineering,* 63(1):155–166, Oct.

GOEDE, R. & KRUGER, H.A. 2004. Evaluating information security awareness - a comparison of different research paradigms. Research Report No FABWI-N-RKW:2004-135. Potchefstroom: North-West University.

HAMPEL, F. 2002. Some thoughts about classification. (*In* Jajuga, K., Sokolowski, A. & Bock, H.H., *eds.* Classification, clustering and data analysis: recent advances and applications: papers presented at the eighth Conference of the International Federation of Classification Societies, held in Cracow Poland 16 to 19 July, 2002. Germany. p. 5-27.)

HAN, J. & KAMBER, M. 2006. Data mining: concepts and techniques. 2nd ed. San Francisco: Kaufmann.

HORMOZI, A.M. & GILES, S. 2004. Data mining: a competitive weapon for banking and retail industries. *Information systems management*, 21(2):62-71, Spring.

JAIN, A.K., MURTY, M.N. & FLYNN, P.J. 1999. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264-323, Sep.

KARAMI, A., ROWLEY, J. & ANALOUI, F. 2006. Research and knowledge building in management studies: an analysis of methodological preferences. *International journal of management*, 23(1):43-52. Mar.

KIM, J.W., LEE, B.H., SHAW, M.J., CHANG, H. & NELSON, M. 2001. Application of decision-tree induction techniques to personalized advertisements on internet store fronts. *International journal of electronic commerce*, 5(3):45-62, Spring.

KIM, K-J. & AHN, H. 2008. A recommender system using GA K-means clustering in an online shopping market. *Expert systems with applications,* 34(2):1200–1209, Feb.

KNOWLEDGE FACTORY. 2005. ClusterPlus: helping you to better understand and locate your target customers and markets through geo-demographic segmentation. http://www.knowledgefactory.co.za/download.php?view.10 Date of access: 21 Aug. 2007.

KONONENKO, I. & KUKAR, M. 2007. Machine learning and data mining: introduction to principles and algorithms. Chichester: Horwood.

KUMAR, P.R. & RAVI, V. 2007. Bankruptcy prediction in banks and firms via statistical and intelligent techniques: a review. *European journal of operational research,* 180(1):1–28, Jul.

KUNENE, K.N. & WEISTROFFER, H.R. 2008. An approach for predicting and describing patient outcome using multicriteria decision analysis and decision rules. *European journal of operational research*, 185(3):984–997, Mar.

LEE, A.S. 1991. Integrating positivist and interpretive approaches to organizational research. *Organization science,* 2(4):342-365, Nov.

LIAO, S-H. & WEN, C-H. 2007. Artificial neural networks classification and clustering of methodologies and applications – literature analysis from 1995 to 2005. *Expert systems with applications*, 32(1):1–11, Jan.

LIN, K-S. & CHIEN, C-F. 2009. Cluster analysis of genome-wide expression data for feature extraction. *Expert systems with applications*, 36(2):3327–3335, Mar.

LU, C-L. & CHEN, T-C. 2009. A study of applying data mining approach to the information disclosure for Taiwan's stock market investors. *Expert systems with applications,* 36(2):3536-3542, Mar.

MACQUEEN, J. 1967. Some methods for classification and analysis of multivariate observations. (*In* Le Cam, L.M. & Neyman, J., *eds*. Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability, 1966, Berkeley University. p. 281-297.)

MAREE, K., *ed.* 2007. First steps in research. Pretoria: Van Schaik.

MILLS, R. 2000. Developments in cost and management accounting. Accounting and finance, cost and management accounting. Nov 2000.

MINGOTI, S.A. & LIMA, J.O. 2006. Comparing SOM neural network with Fuzzy c-means, K-means and traditional hierarchical clustering algorithms, *European journal of operational research,* 174(3):1742–1759, Nov.

MOSTAFA, M.M. 2009. Modelling the efficiency of top Arab banks: a DEA-neural network approach. *Expert systems with applications,* 36(1):309–320, Jan.

MURPHY, J., SCHEGG, R. & OLARU, D. 2007. Quality clusters: dimensions of email responses by luxury hotels. *International journal of hospitality management*, 26(3):743-747, Sep.

NGAI, E.W.T., XIU, L. & CHAU, D.C.K. 2009. Application of data mining techniques in customer relationship management: a literature review and classification. *Expert systems with applications*, 36(2):2592–2602, Mar.

O'DONNELL, C.J., VAN DER WESTHUIZEN, G. 2002. Regional comparisons of banking performance in South Africa. *The South African journal of economics*, 70(3):485-518, Mar.

PORCEL, J.M., ALEMÁN C., BIELSA, S., SARRAPIO, J., DE SEVILLA, T.F. & ESQUERDA, A. 2008. A decision tree for differentiating tuberculous from malignant pleural effusions. *Respiratory medicine,* 102(8):1159-1164, Aug.

PRASAD, P. 2005. Crafting qualitative research: working in the postpositivist traditions. Armonk, NY: M.E. Sharpe.

PREMACHANDRA, I.M., BHABRA, G.S. & SUEYOSHI, T. 2009. DEA as a tool for bankruptcy assessment: a comparative study with logistic regression technique. *European journal of operational research,* 193(2):412–424, Mar.

PYLE, D. 1999. Data preparation for data mining. San Francisco: Kaufmann.

RAMAKRISHNAN, R. & GEHRKE, J. 2003. Database management systems. 3rd ed. Boston: McGraw Hill.

RENDER, B. & STAIR, R.M. 2000. Quantitative analysis for management. 7th ed. Upper Saddle River, NJ: Prentice Hall.

SAMOILENKO, S. & OSEI-BRYSON, K. 2008. Increasing the discriminatory power of DEA in the presence of sample heterogeneity with cluster analysis and decision trees. *Expert systems with applications,* 34(2):1568-1581, Feb.

SAS INSTITUTE. 2008. Prediction & prevention. http://www.sas.com/success/absa.html Date of access: 14 Jan 2008.

SAUNDERS, M., LEWIS, P. & THORNHILL, A. 2000. Research methods for business students. 2nd ed. London: Prentice Hall.

SEOL, H., CHOI, J., PARK, G. & PARK, Y. 2007. A framework for benchmarking service process using data envelopment analysis and decision tree. *Expert systems with applications,* 32(2):432–440, Feb.

SEWITCH, M.J., LEFFONDRÉ, K. & DOBKIN, P.L. 2004. Clustering patients according to health perceptions: relationships to psychosocial characteristics and medication nonadherence. *Journal of psychosomatic research*, 56(3):323-332, Mar.

SEXTON, T.R., SLEEPER, R. & TAGGART, R.E. 1994. Improving pupil transportation in North Carolina. *Interfaces*, 24(1):87-103, Jan.

SHERMAN, H.D. & GOLD, F. 1985. Bank branch operating efficiency: evaluation with data envelopment analysis. *Journal of banking and finance*, 9(2):297-315, Jun.

SHERMAN, H.D. & LADINO, G. 1995. Managing bank productivity using data envelopment analysis. *Interfaces*, 24(2):60-70, Mar.–Apr.

SINKA, M.P. & CORNE, D.W. 2005. The BankSearch web document dataset investigating unsupervised clustering and category similarity. *Journal of network and computer applications*, 28(2):129–146, Apr.

SOHN, S.Y. & MOON, T.H. 2004 Decision tree based on data envelopment analysis for effective technology commercialization. *Expert systems with applications,* 26(2):279-284, Feb.

STEFURAK, T. & CALHOUN, G.B. 2007. Subtypes of female juvenile offenders: a cluster analysis of the million adolescent clinical inventory. *International journal of law and psychiatry*, 30(2):95–111, Mar.-Apr.

SUBRAMANIAN, R., SCHEFF, R.P. & QUILLINAN, J.D. 1994. Coldstart: fleet assignment at Delta Airlines, *Interfaces*, 24(1):104-120, Jan.

SUN, J. & LI, H. 2008. Data mining method for listed companies' financial distress prediction. *Knowledge-based systems*, 21(1):1–5, Feb.

VASSILOGLOU, M, & GIOKAS, D. 1990. A study of the relative efficiency of bank branches: an application of data envelopment analysis. *Journal of the Operational Research Society*, 41(7):591-597, Jul.

WALSH, S. 2005. Applying data mining techniques using SAS enterprise miner course notes. Johannesburg: SAS Institute Inc. (Course code, ADMT5 / Book code, 59758)

WANG, W. & ZAÏANE, O.R. 2002. Clustering web sessions by sequence alignment. (*In Proceedings of the 13th International Workshop on Database and Expert Systems Applications.* DEXA, 2-6 September 2002. p. 394-398.)

WATKINS, A., HUFNAGEL, E.M., BERNDT, D. & JOHNSON, L. 2006. Using genetic algorithms and decision tree induction to classify software failures. *International journal of software engineering and knowledge engineering,* 16(2):269-291, Apr.

WHITTEN, J.L., BENTLEY, L.D., DITTMAN, K.C. 2001. Systems analysis and design methods. 5th ed. Boston, MA: Irwin/McGraw-Hill.

ZHAO, H., SINHA, A.P. & GE, W. 2009. Effects of feature construction on classification performance: an empirical study in bank failure prediction. *Expert systems with applications,* 36(2):2633–2644, Mar.

ZHAO, Y. & ZHANG, Y. 2008. Comparison of decision tree methods for finding active objects. *Advances in space research,* 41(12):1955–1959.

ZAÏANE, O.R, FOSS, A., LEE, C-H. & WANG, W. 2002. On data clustering analysis: scalability constraints and validation. (*In* Chen, M-S., Yu, P.S. & Liu, B., *eds.* Advances in knowledge discovery and data mining: 6th Pacific-Asia Conference (PAKDD) 2002 Taipei, Taiwan. 6–8 May 2002: proceedings. p .28-39.)