UNIVERSITY OF NAIROBI

Use of GIS and Association Rule Mining in Guiding Strategic Business Expansion Planning

Case Study: Chase Bank (K) Ltd

BY

LABAN KIPROTICH RONOH

F56/P/71318/2007

JULY 2009

Use of GIS and Association Rule Mining in Guiding Strategic Business Expansion Planning

Case Study: Chase Bank (K) Ltd

BY

Laban Kiprotich Ronoh

F56/P/71318/2007

A Project Report Submitted in Partial Fulfillment of The Requirements For The Degree of Master of Science in Geographic Information Systems (M.Sc. in GIS) in the University of Nairobi.

JULY 2009

DECLARATION

I, the undersigned, declare that this is my original work and has not been submitted to any College, Institution or University other than the University of Nairobi for academic credit. All sources of information have been specifically acknowledged.

SIGNED ____________________ DATE 6/08/2009

Laban Kiprotich Ronoh

This project has been presented for examination with my approval as the appointed supervisor.

SIGNED ____________________ DATE 07.08.2009

Dr.-Ing. J. B. K. Kiema

DEDICATION

I would like to dedicate this project to the following:

Mrs. Enid Rael Ronoh my wife, for the understanding throughout the study period, Ryan and Clifford our children for all their moral support and inspiration given. These young guys often asked when daddy would complete school so that he could create time to take them out!

Special dedication goes to my parents, for imparting in me the culture of patience, diligence and discipline, without which, I would not have been able to undertake this postgraduate course.

Also my dedication goes to Parmain ole Narikae and Zachary Maritim for their mentorship, support and inspiration.

ACKNOWLEDGEMENTS

The work presented here would have not been possible without the assistant of various people and institutions to whom I am utmost grateful.

Dr. J. B. K. Kiema my supervisor, whose guidance, criticism, suggestions and encouragement enabled me to complete this project.

Special thanks also goes to Paul Yego for the assistance and support in data collection and analysis he extended to me to make this project a success.

Finally to those who assisted me in collecting the data and more so to Chase Bank's personnel including Ian Kingara (*Head of Operations*), John Mathiaka (*Head of ICT*), Makarios Agumbi (*Head of Finance*) and Steve Simiyu (*Systems Administrator*) who willfully devoted their valuable time in giving out information without which this work would not have been completed.

ABSTRACT

In recent years, optimal site selection has become one of the main concerns for managers of business enterprises. In addition, various kinds of spatial and non-spatial parameters influence the efficiency of new branches. These factors have a direct relation with site selection indicators. In this research project, the use of Geographic Information Systems (GIS) and Data Mining (DM) to determine and extract useful knowledge not only to help managers make better decisions for site selection, but also for extracting associations between selected parameters is investigated. The study also attempted to find a link between a mathematically determined efficiency measure and spatial/general association rules, which is a database method in data mining. During the research, the study area was classified into three different classes as 'high', 'average', and 'low' according to the efficiency and turnover measures. Afterwards, in each class an *a priori* like algorithm was used to establish the most frequent item sets and predict an average range of efficiency. In general, as the efficiency measure in the low class had a higher frequency than in other classes, negative rules were obtained rather than positive rules. In addition, the association rules for the small scale gave more meaningful results than those of the large scale. The reason was in the use of real parameters instead of aggregated parameters. The usability of this method was not absolutely good with this data set and it is recommended that normal distributed efficiency measure data be employed to find association rules in all the classes. Finally, for the site selection issues, the managers can use this method as a comparison factor, among different candidate areas. They can rely on the validation measures such as support, confidence, lift and leverage to select the best location for a new site.

Keywords:

Geographic Information Systems, Data Mining, Trade Area, Spatial Association Rule, Efficiency.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

1.1 Background to the Study

The role of information and communication technology (ICT) has grown and changed continuously in the banking sector. The banking industry has used ICT to increase volume of transactions as well as development of new products. ICT applications have ranged from back-office processing; mortgage and loan application processing, and the electronic funds transfer to more strategic innovations such as automated teller machines and virtual banking services like mobile banking and money transfer services. The use of ICT has also had some important customer – supplier effects. For the customers of service providers, it has been used to improve the quality and variety of services in many industries, especially through its ability to amass, analyze, and control large quantities of specialized data [National Research Council (NRC), 1994]. Such improvements include error reduction or increased precision, faster or more convenient service, and improved security, safety, and reliability.

According to NRC (1994), the banking industry is a major factor in the economy. Although it has grown at a moderate rate over the last two decades, the most significant changes in this period concern its character rather than its size. For most of its history, banking has been subject to extensive state regulation. However, partial bank deregulation in the late 1970s and early 1980s led to a sharp increase in the variety of services and products offered by commercial banks. Driven by both technology and competition from non-bank financial institutions, increasing product diversification continues today in commercial banking, although it is still constrained to some extent by current regulations.

Given the magnitude of the banking industry's investments in ICT over the last two decades, large increases in productivity might have been expected. One reason these have not appeared in measures of productivity is that such measures in the banking industry remain highly problematic. For example, the

24-hour availability through automated teller machines of many deposit and withdrawal services previously accessible only during bank hours. Another reason for the lack of large increases in measured productivity is that early applications of ICT proved to be costly and cumbersome. Software and equipment had to be updated and replaced frequently. A great array of new products constantly called for new software and communication capabilities. Cost control and productivity tracking systems lagged behind the new technologies in a rapidly changing marketplace. The result was that tangible paybacks from ICT investment were delayed (NRC, 1994).

The development of a modern banking technology began in the 1960s. It was during this time, when computer and microelectronics sectors started their growth (Freeman and Perez, 1988). Computers made it possible to handle a huge amount of transactions in a very short time. These new opportunities and changes had an important effect on the organisation of work; banking personnel left routine based and time consuming work to computers and began to concentrate on the service-sector. This was beginning of a new paradigm, "the information communication technology paradigm", and banking was probably the first major service branch which adopted new information technologies extensively (de Wit, 1990).

To bring services closer to a customer and to guarantee the opportunity to use them anytime a customer wants to have been the most important target in banking during the last twenty years. The continuing development of more and more complicated back-office systems would not have been possible without information and communications technology. In many cases, computers have replaced banking personnel and they have become the most important factor behind the decreasing amount of working places. This new information technology led to savings in labour costs, but it also originated a process of saving in other categories of capital as well, like buildings (de Wit, 1990).

According to de Wit (1990), a bank office would be more technology based. He further noted that a bank office in the future is going to look like a department store, where customers can make their daily "purchases" with help of machines. The personnel would be needed to make the most complicated tasks and to give some advice and information to customers. From the author's point of view, what *de Wit* visioned over nineteen years ago, has now become almost a reality. The machines have replaced the service counters and the personnel are walking around the shop helping customers to use these highly developed machines.

In the 1960s, the increasing amount of transactions in the banking sector created a growing demand for new personnel. This increasing amount of personnel was one of the main reasons that forced banks to use the automatic data processing. The competitiveness was developed by improving the effectiveness of business. There were two main aspects, which played an important role in this process: one was the requirement of personnel and the other was the cost-savings which became possible by simplifying the routine-based work. Also by computerizing these basic routines it became possible to develop some new types of services. One of the first services was so called "wages and salaries direct to bank accounts", which helped people to familiarize with bank services in their daily businesses. This new service led to a rapid growth of banking customers (Ibid, 1990).

The 1970's was a decade, when banks changed-over to the real-time data processing systems. A large percentage of paper-based transactions were transmitted and processed electronically. Automated Teller Machine services and direct electronic deposits and withdrawals by large automated users replaced many paper processes. As new products and services expanded, and as margins became less predictable, commercial banks began investing in front-office automation to provide better information to personnel related to customer service and to enhance the delivery of products and services. The amount of transactions increased much faster than was expected, so the real-time system was the only alternative to keep the amount of personnel constant and to hold down the increasing costs of handling information and the operating rooms (Ibid, 1990).

Banking services for private customers largely consist of personal service, supplying bank notes and providing payment services. In the 1980's, it became possible to serve customers outside the bank-office in the form of self-service. This kind of self-service has provided a new and a flexible way for customers to conduct their banking affairs. The reason for this fast growing self-service have been short opening hours at bank offices, rush hours at specific times, the huge amount of withdrawals (about 50% of all transactions) and also an attempt to keep down the rising costs. The first step in self-service was the introduction of Automated Teller Machines – ATMs (Ibid, 1990).

In the late 1980s, the banking industry began to focus on automation of data communications. The installation of on-line terminals in the early 1970s enabled automation of the customer interface and front office applications in such areas as corporate treasury. ATMs, first introduced in the late 1970s in other parts of the world and early 1990 in Kenya, have become an agent of a strategic change in banking.

The last ten years of 20^{th} century have been characterized by a rapid growth of personal computers. Personal computers have become very common especially in the 21^{st} century and have created a demand for new types of innovative banking services. Home-based banking as well as internet banking by using the home computer is the newest service for banking customers and all these innovations have been necessitated by the improvements in technology. Connectivity via broadband and fibre optics will therefore continue to shape the introduction of innovative banking products in the ever competitive banking environment.

Looking at the general banking landscape therefore, the distinguishing factor amongst banks is basically the quality of service they offer. With the intense competition and the ever growing customer sophistication, banks will have no choice but to improve the quality of services rendered. In recent years, and as detailed in the background information above, significant advancement in ICT has accelerated and broadened the dissemination of financial information and services and also increased complexity. Banks have also continued in their aggressive marketing, although with a large portion of the market terrain yet uncovered. Indeed, the challenge is enormous.

As detailed above, banking services is data driven. This is where Geographic Information System (GIS) becomes very useful. GIS customized to corporate requirements, holds immense potential for the productivity of any organization. GIS technology has evolved into a formidable tool by through which the corporate world can use spatial information to manage their businesses. GIS also allow the users to spatially visualize data thus revealing hidden relationships, patterns and trends.

It offers a platform for developing a customer-centric business model and an integrated environment to help banks in decision making, strategic planning, effective resource management and operations management. This will obviously boost customer satisfaction, stimulate business growth and engender customer base expansion.

Expansion strategies in the banking industry are driven largely by the desire to have a presence in a particular locality and neighbor-hood that is considered to represent the niche market for the product range and services provided. Feasibility studies reports are therefore derived based on the performance history attributed to existing branches as a model of prediction.

Site selection for banks is important for business enterprise development. In practice, so much money is wasted because of lack of knowledge and objective strategies for finding new sites. Integration of GIS technology and Data Mining (DM) provides a scientific alternative and objective means of determining the best locations for the establishment of delivery channels. Optimal site selection will depend on a number of parameters that will have spatial, social, technical and economical characteristics which are either quantitative or qualitative.

1.2 Problem Statement

Business needs are continually driving the demands for increased capabilities of ICT. In turn, increasingly advanced ICT is being utilized in more and more sophisticated ways by businesses to outdo competition. ICT, which is being deployed as a solution to the increased complexity and uncertainty of the environment, has paradoxically contributed to the situation by "compressing time and distance." In the absence of the present day advances in ICT, would one be talking of globalization or time-based competition? Perhaps not! The pace of complexity is increasing fast. Hopefully, the advances in technology and spatial parameters in location of business outlets as driven by the use of GIS would be able to keep up with the environmental changes and take banking competition to another level.

To survive in the fast-changing environment the "adaptive organization" would be more like a shifting "configuration". Effective implementation of ICT and GIS would decrease vulnerability by reducing the cost of expected failures and enhance adaptability by reducing the cost of adjustment. Overall, the impact of the current technology investment boom in the financial services sector is difficult to assess. Productivity in financial services, like productivity in the rest of the service sector, is very hard to measure. The problem is due partly to the difficulty of measuring output accurately when the quality of service is changing as a result of such factors as greater convenience, optimal outlet location, speed and lower risk of doing business.

Moreover, a number of parameters are involved while making strategic planning decisions in banks. An integrated GIS/Data Mining model can be used to evaluate "what if" scenarios by using interrelationships between land use factors, infrastructure capacities and economic growth among other parameters. This will no doubt be very useful to management as it offers a good platform for major decisions to be thoroughly evaluated before they are executed.

Expansion planning strategy requires the modeling of spatially relevant data and offering fast and cost effective site analysis to effectively select a new site. With GIS, one can choose suitable site for new branch/delivery channel by using a combination of population density, land/building availability, costs and availability of infrastructure, crime rate analysis among other factors. These parameters can readily be integrated in a GIS and coupled with its ability to display these features pictorially on the map, can aid analysts in deciding if a site meets the specified criteria. In addition, the demographic content of GIS can also aid in making decisions such as the maximum number of branches a region can support.

This research project aimed at modeling an integrated approach of using GIS and Data Mining (DM) to provide a scientific and more objective means of aligning a bank's expansion strategy with spatial parameters that determine the relationships of a given location and the productivity/efficiency parameters, while taking cognizance of the competitor environment. It aimed at developing a method for discovery of different spatial and non-spatial parameters that have an effect on site selection for bank branches.

The innovation of coupling ICT innovations together with GIS will enable the modeling of a spatial method to help managers in allocating optimal new branches/ATM outlets based on the identified parameters that link the efficiency of new branches from existing branch efficiencies using spatial association rules derived from a pool of parameters before site selection is carried out. Parameters taken into consideration include profit, income and turnover to augment the measurement of productivity and prediction of efficiency for both the existing and the proposed new branches within the area of study.

1.3 Objectives of the Study

The general objective of the research project was to evaluate the spatial relationships that can be used to guide strategic expansion planning as applied to banking environment by employing the use of innovations provided by ICT, and GIS and Data Mining environment in the banking industry. The study aimed at analyzing and reviewing how the new developments in information technology can influence/affect strategic outlet expansion planning in rolling out of banking services and methods as well as its implications for the bank customer base.

The specific study objectives of the research were:

1. To evaluate the efficiency of the existing branches, derive spatial association rules that can be used to predict the efficiency of new delivery channels to be able to determine the best sites for business expansion.
2. To determine the spatial and non-spatial parameters that influence the target parameters positively and negatively – association rules for optimization, total turnover and efficiency.
3. To carry out competitor analysis in respect of the players in similar scale of operation and niche market and display the distribution in form of a digital map.
4. Recommend an implementation plan/road map of an integrated GIS/Data Mining Model as an expansion strategy decision support system and tool in the banking industry.

1.4 Research Questions

The study was guided by the following research questions:

(a) What theory of spatial parameters and rules is best applied for the problem statement?

(b) What are the spatial and non-spatial parameters that both positively and negatively influence determination of suitable location for banking outlets?

(c) How have the banks responded to the innovations offered by the changing information communication technology environment?

(d) How are such developments and innovations of information communication technology like GIS and Data Mining implemented for the case of a bank?

1.5 Importance and Justification of the Research

The results of the study would help management of banks to appreciate the need of information technology in the changing environment, more so the application of the recent technological innovation of GIS and Data Mining and try to invest in these cutting edge technology in order to compete effectively in the industry.

In addition the research is expected to come up with suitable suggestions in terms of strategic expansion planning through the integration of GIS and Data Mining, specifically the spatial/locations rules to model guiding principles for expansion planning and target marketing.

1.6 Scope of the Study

The study focused on the use of scientific and objective methods to guide strategic expansion planning. Specifically, through the use of GIS and Data Mining one can derive Spatial Association Rules that form a basis for decision making and help managers in formulation of competitive expansion strategies in the banking sector. The sector is characterized by changing customer needs, industry trends and stiff competition, hence the need to analyze the importance of information technology in strategic decision making. The study was intended to cover all the commercial banks in Kenya but due to geographical location, time and the nature of the research (academic), only a selected number of banks within Nairobi City have been sampled. In particular, and for financial analysis, Chase Bank (K) Ltd has been used as a case study. Overall, both spatial and financial productivity/efficiency data were analyzed and the results generalized to represent a 'prediction model' for the entire sector.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

This chapter focuses on works of other writers regarding the subject topic of study. Areas of interest include history of information technology, general developments in information technology, information technology as an innovative strategy, the importance of information technology in banking, GIS and Data Mining technology and applications in the business environments of information technology and the use of information technology in the regulation of financial institutions. Indeed, GIS and Data Mining as well as Decision Support Systems are products of innovation that have been made possible by the advancement of Information and Communication Technology (ICT).

The concepts discussed in the literature review contains other major disciplines like Spatial Decision Support Systems, Data Mining and the different methods and specially association rule method as the main and essential concept in the research. In addition, the relation between the spatial data and the association rules are described briefly.

Finally, *DEA* as a kind of mathematical model is covered as well as its application in measuring the efficiency of any financial institution. Other related fields of science such as spatial economics will be discussed. The topics discussed come from different branches of science but this research will try to combine them. Finally, the term '*DEA* based spatial association rule mining' will be the innovation of our research in the scientific point of view.

2.2 Business Environment of ICT

The world of banking differs quite a lot in different countries. To understand the differences between countries, Rosenberg (1994) has found out some "needs" which play different roles in different countries. Differences in the resource endowment and demand conditions of an economy are showing the way as well as the kind of inventions that will be profitable to develop and exploit. Each country has its own visions about what is important and what might be worth developing just in that specific country. Rosenberg (1994) argues, that only those inventions, which are compatible with a country's needs will be successful. At different times, there are some innovations which are easier to create than others, and there might be some great differences with other countries. The level of technical knowledge, as well as economic forces tends to push economies in different directions.

In a survey article in *The Economist*, John Browning (1990) wrote: "Information communication technology is no longer a business resource; it is the business environment." His statement is not far from truth. Ongoing advances in information and communication technology (ICT), along with increasing global competition, are adding complexity and uncertainty of several orders of magnitude to the organizational environment. One of the most widely discussed areas in recent business literature is that of new organizational network structures that supposedly hold the promise of survival and growth in an environment of ever-increasing complexity. How can ICT help the organizations in responding to the challenges of an increasingly complex and uncertain environment? How can ICT help the organizations achieve the "flexible" organization structure? The answers to these questions lie in the increasing scope of innovations derived from the changing ICT environment. Increase of commercial off-the-shelf applications and increased understanding of customizable languages and packages will no doubt revolutionize the application of ICT innovations in the banking industry.

Moreover, the increased competition for customer convenience will accelerate adoption of scientific models of optimization, all geared at increasing the range of products available to the informed customers who are targeted by most banks to shore up the productivity and efficiency levels of the business outlets and expansion outreach.

2.3 Financial Institutions Regulation and ICT

Regulated financial intermediaries can only stabilize the operation of global markets if they retain a significant share of those markets. However, regulated financial intermediaries are hobbled by burdensome regulations designed to protect the public. When they have few alternatives to regulated financial services organizations, they may be powerless to stop the erosion of their market share in favor of unregulated competitors whose operating costs are lowered due to fewer costs of compliance with regulations. Regulators face a difficult task in deciding whether to lighten the regulatory burdens now imposed on financial institutions to permit them to compete more equally with non-banks or whether to try to regulate the new entrants to the marketplace. The climate of deregulation in Kenya and elsewhere would make it difficult to create new regulations to govern innovative business trends that have been made possible by advances in computing technology. Virtual banking services offered by mobile telecommunication industry is indeed one such unregulated financial services industry that poses a new dimension to banking and financial services competition.

Regulators may have few alternatives to using the greater access provided by global networked computer systems as an important tool in working to make markets operate more efficiently and safely. The same access enjoyed by market participants would permit regulators to disseminate information that market participants' need to make sensible decisions and to protect themselves. If markets do actually become more efficient, then many of the traditional bases for regulatory activity may diminish (Kenya web archives, 2002).

The regulator will no doubt be required to innovate other ways and means of regulating and monitoring the financial institutions in the wake of increased application of innovative ICT initiatives to drive business. Feasibility studies will increasingly become scientifically oriented as banks outwit one another in the competitive environment. The vetting and approval of such feasibility studies will require the financial services regulator – Central Bank of Kenya to employ innovative scientific measures in the analysis, vetting and approval process for banking outlets among the over 40 banks in the country.

2.4 Trends in Business GIS

Business use of GIS covers a spectrum of GIS applications. The use of GIS applications is still somewhat fragmented and there is need for further integration with other forms of ICT. The trend in ICT applications has been for initial operational use, followed in turn by sophisticated specialist decision making and executive management applications (Nolan, 1973, 1979). The model described in this literature review is the use of use of an organization's level of expenditure in ICT to compute growth. A more recent approach comprises the steps of *initiation, expansion/contagion, formulation/control, integration, data administration and maturity* as described by J. Pick, (2005).

The GIS community is generally attempting to integrate multiple spatial data sources and new metadata standards among other initiatives with the aim of facilitating business GIS applications (Goodchild *et al*, 2003).

Indeed, with data organization made easier through the use of spatial databases, business users will be able to take advantage of better integrated GIS data to extend the area of business where GIS is used. As earlier noted, ICT adoption by banks to drive business applications is an important factor in driving innovation and use of non-customary business technology informatics. Knowledge of these applications will to a greater extend help in popularizing and creating an appreciation of the technology. Top business managers coupled with this knowledge will be able to sufficiently drive the enthusiasm required to integrate spatial decision support in business decisions.

2.5 Spatial Decision Support Systems

The spatial decision support systems have been extensively and adequately covered in the literature - Craig and Moyer, (1991), Densham, (1991), Goodchild and Densham, (1990), Moon, (1992), NCGIA, (1992). The need for using such systems comes from situations in which complex spatial problems are ill- or semi-structured and decision makers cannot define their problem or fully articulate their objectives. The decision making process adopted to solve semi-structured spatial problems is often being perceived as unsatisfactory by decision makers. What they really need is a flexible, problem-solving environment in which the problem can be explored, understood and redefined, trade-offs between conflicting objectives investigated and priority actions set.

Densham (1991) quotes Geoffrion's (1983) definition of Decision Support Systems suggesting that DSS has six characteristics:

1) explicit design to solve ill-structured problems;
2) powerful and easy-to-use user interface;
3) ability to flexibly combine analytical models with data;
4) ability to explore the solution space by building alternatives;
5) capability of supporting a variety of decision-making styles; and
6) allowing interactive and recursive problem-solving.

He then adds to the list the distinguishing capabilities and functions of Spatial Decision Support Systems, which need to:

1) provide mechanisms for the input of spatial data;
2) allow representation of the spatial relations and structures;
3) include the analytical techniques of spatial and geographical analysis, and
4) provide output in a variety of spatial forms, including maps.

As an extension of DSS, SDSS is a computer-based information system used to support decision-making where it is not possible for an automated system to perform the entire decision process. The intangible factors in the decision making process may be accounted for through information supplied and choices made by a decision maker who operates the SDSS interactively or operates it through an analyst. The above suggest that spatial decision support systems may be developed as general purpose tools for decision making (Onsrud's paper in Goodchild and Densham, 1990).

Just like DSS, SDSS have four modules: a data management system, analytical modeling capabilities and analysis procedures, display and report generators, and a user interface. Densham (1991) separates the display generator and the report generator into two modules and describes the user interface as a module encompassing the other four modules. He also highlights the generation and evaluation alternatives procedure in this interactive, iterative, and participatory process.

Also like DSS, SDSS have three levels of technology:

1) an SDSS toolbox, i.e. a set of hardware and software components that can be assembled to build a variety of system modules;
2) an SDSS generator, i.e. hardware and software modules that can be assembled to build a specific SDSS, and
3) specific SDSS (Sprague, quoted in Densham, 1991).

Densham (1991) also distinguishes five functional roles:

1) the SDSS toolsmith develops new tools for the SDSS toolbox;
2) the technical supporter adds components to the SDSS generator;
3) the SDSS builder assembles modules into specific SDSS;
4) the intermediary sits at a console and interacts physically with the system;
5) the decision maker is responsible for developing, implementing and managing the adopted solution.

Armstrong, (1990) looks at the expert analyst required to operate the system as posing a barrier to decision-makers who must translate the problem into a form that can be understood by experts who, in turn must translate their understanding of the problem into a form that can be modeled by software. This prevents decision-makers from directly interacting with the problem and may prevent them from discovering how intermediate decisions affect final outcomes.

2.6 Users of Spatial Decision Support Systems

As Cooke (1990) puts it: "A decision-maker's job is to make decisions, not deal with the technical minutia surrounding geographic databases..." Nothing could be more horrifying to a decision-maker than seeing his/her problem dealt with by an expert system, leaving him/her virtually in the spectator's seat. For decision support systems to have profound impacts on managers' activities, they should be integrated into an organization's decision-making culture and process. Beaumont (1990) appreciates that current usage of DSS is predominantly by "middle" management rather than top management.

Who are the potential SDSS users anyway? The decision makers? The intermediary? Other actors and players? To enhance organizational efficiency and effectiveness, support must be developed for group workings since discussions, negotiations, bargaining with colleagues are important dimensions of decision making. While users of SDSS can be individual or group decision-makers, technical advisers, planners, interest groups and "the public" at large, many authors seem to consider the litmus test for such systems to be their ability of addressing the immediate needs of top decision makers. Gould (1990) wants to see SDSS designed for users who are themselves decision-makers.

Once the target user has been identified, the difficulties and barriers to the widespread adoption and use of SDSS seem to compound. Most systems builders seem to be unaware of the complex nature of the decision-maker's job and of the assumptions and "hidden agendas" that "cloud" the "rational" process of decision. Winograd and Flores (1986) discuss some of the dangers that potentially attend the use of decision support systems, such as:

1) orientation to choosing;

2) assumption of relevance, i.e. the assumption that the things the installed computer system does are the ones most relevant to the decision-maker;
3) unintended transfer of power to programmers, software designers and analysts;
4) unanticipated effects, desirable and undesirable;
5) obscuring responsibility in interpreting the machine as making commitments;
6) false belief in objectivity.

A good SDSS seems to have to deal with the capabilities of humans as problem solvers, with short term and long term limitations, associative memory structures, conservatism biases, and decision-making illusions. And it must leave a maneuvering room to the user in order to have a chance to be accepted. Technically, this requirement imposes the incorporation of the user's judgments, values and knowledge in the decision support system.

2.7 Data Mining

Knowledge improvement has led scientists to think about analysis and extraction of useful information from large databases. Previously, researchers tried to improve understanding with methods and techniques, such as statistical analysis and various mathematical models. Due to the increase of database transactions in large organizations and specifically in governmental institutes, unstructured analysis became one of the main challenges in such organizations.

In the middle 1990's, an important revolution happened in the field of knowledge discovery in databases. The foundation of data mining was based on statistical methods and gradual improvement of different research works caused many developments in advanced use of large databases. In general, there are multiple definitions for data mining as follows:

Data mining: Simply stated, data mining refers to extracting or mining knowledge from large amounts of data (Han *et al*, 2006).

This definition is a good starting point but if one wants to define the data mining concept the following definition is more precise:

Data mining: The extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amounts of data (Hand *et al*, 2001).

There are more slightly different definitions in the literature such as the following:

Data mining: The analysis of (often large) observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner (Han *et al*, 2006).

2.8 Data Mining Methods

Generally, any kind of useful knowledge extraction from a data set with some statistical or query-based method is the result of a simple data mining. There exist various types of algorithms used in the classic data mining. These methods are categorized as follows:

1) Classification
2) Estimation
3) Prediction
4) Clustering
5) Association

The descriptions in this area are extracted from (Larose, 2005).

2.8.1 Classification

In classification methods, usually there is a categorical target variable with which all the data are categorized. In other words, the data mining model tries to examine a large data set of records both with the target variable and other fields as input. For example, suppose we have a data set about annual income of the employees with their age, gender and occupation. In this example, the target variable is income and it can, for instance, be categorized into three different ranges as follows: 'High', 'Average', 'Low'. Here, the predictors would be age, occupation and gender, from which using the data mining engine, three classes will be generated.

The next step is the training of the model, after which any new object can be classified in a specific class. There are different types of mapping classification methods as discussed below. The contents are extracted from Campel (2001) and Krygier *et al*, (2005).

- **Natural Breaks:** This method is a data classification method that divides data into classes based on the natural groups in the data distribution. It uses a statistical formula (Jenks optimization) that calculates groupings of data values based on data distribution, and also seeks to reduce variance within groups and maximize variance between groups. Natural Breaks method is based on subjective decision and it is best chosen for combining similar values in such a way that there is no extreme value with high tolerance in a class.

- **Quantile:** The quantile classification method distributes a set of values into groups that contain an (approximately) equal number of values. This method attempts to place the same number of data values in each class and will never produce empty classes or classes with few or too many values.

- **Equal Interval:** The equal Interval Classification method divides a set of attribute values into groups that contain an equal range of values. This method works better with continuous set of data because the map designed by using equal interval classification is easy to accomplish and read. However, performs badly with clustered data because many items may wind up in just one or two classes while others will have no features at all.

- **Standard Deviation:** The standard deviation classification method determines the mean value, and then places class breaks above and below the mean at distances of either 0.25σ, 0.5σ or so, until every data value is contained within a class. By σ we mean the value set is standard deviation. Values that are beyond a threshold distance from the mean are usually aggregated into two outlier classes: small and large, for instance.

2.8.2 Estimation

Estimation enables us to obtain a parameter estimate from the existing data. In this area, regression is a commonly used technique. It results in a formula with which new data can be assigned as an estimate for the parameter.

Using one of the regression methods, the relationship between one or more response variables (also called dependent variables, explained variables, predicted variables, or regressands, usually named Y), and the predictors will be estimated. For example, a manager of an institute wants to know the total budget for next year with respect to the number of employees and existing customers. Considering the previous existing parameters and also total budget, a mathematical formula in the form of $Y = f(x, t, \ldots)$ will be found in which x, t, . . . are the variables, and Y is the total budget estimation.

2.8.3 Prediction

Prediction has similarity with the previous methods of estimation and classification. In addition, for predicting phenomena, different types of method can be used, such as statistical modeling or classification but the point is how much the prediction will be different from the reality. A good example in this research domain is predicting the number of accidents for the next year based on historic data. These kinds of phenomena are independent during time and each year it can increase or decrease.

Sometimes in prediction, we cannot find a very good pattern for some phenomena. This is the main distinction between prediction and previous methods. In addition, the reliability of prediction is less than that of other techniques in data mining because instead of exploring inside the data, future phenomena are considered.

2.8.4 Clustering

A common method in data mining is putting similar objects in a group, which is called clustering. Generally, clustering methods are similar to classification but the difference is that in clustering we do not have target variables such as *'high'*, *'average'* and *'low'*.

Actually, clustering algorithms try to find similarities in the data rather than to make predictions about a target variable. These methods find out maximal sets of homogeneous records in a way that minimizes similarity with other clusters. A good example of this method is in fraud detection for the banking industry. In this case, the responsible manager wants to know different costumer behavior segmentations to find unusual bank transaction patterns.

2.8.5 Association

One of the main issues in Association methods is finding relations or connections between attributes of a data set. This method in the business world is sometimes called affinity analysis or market basket analysis (Larose, 2005)

In association methods, an algorithm tries to find rules in the form of 'if antecedent, then consequent'. Such rules must be associated with adequate amounts of support and confidence.

2.9 Association Rule Mining

As mentioned above, association rule mining is one of the most important methods in the data mining concept. The general purpose is to find associations or relationships between item sets.

In data mining terminology, three main definitions are considered. An item corresponds to attribute-value pair, which in this research project is each of the existing parameters. A transaction is a set of items. Each transaction in the set gives us information about which items co-occur in the transaction. A frequent item set is such an item set in which the number of occurrence in the transaction is more than a minimum. In addition, there is a constraint for our work in which we are not allowed to have the same parameter twice in the frequent item set. From the late 1990s, the following theory was developed by Agrawal *et al*, (1993):

> *Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of items. Let D, the task relevant data, be a set of database transactions where each transaction T is a set of items such that $T \subseteq I$. A unique identifier, namely TID, is associated with each transaction. A transaction T is said to contain X, a set of some items in I, if $X \subseteq T$. An association rule implies the form of $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$ and $X \cap Y = \emptyset$.*

2.10 Support and Confidence

In the association rule mining, there are methods for checking the validity of rules. The rule $X \Rightarrow Y$ holds in the transaction set *D* with *confidence c* when *c%* of the transactions in that contain *X* also contain *Y*. The rule has *support s* in the transaction set *D* if *s%* of the transactions in *D* contains $X \cup Y$. The probabilistic formulae 2.1 and 2.2 helps in understanding support and confidence.

$$\text{support, } s = P(X \cap Y) = \frac{\text{\#of transactions containing both X and Y}}{\text{\#of transactions}} \quad (2.1)$$

$$\text{confidence, } c = \frac{P(X \cap Y)}{P(X)} = \frac{\text{\#of transactions containing both X and Y}}{\text{\#of transactions containing X}} \quad (2.2)$$

2.11 *A priori* Algorithm

The *a priori* algorithm is a powerful algorithm for mining regular item sets in the association rule method. It applies the a priori property: Any subset of a frequent item set must be frequent. The background of the algorithm is the use of prior knowledge about frequent item sets already detected. The a priori algorithm uses an iterative concept known as a level-wise search. If we consider k as an arbitrary level, then k-item sets are used to explore (k+1)-item sets. In the beginning, the set of common 1-item sets is found. This set is represented as L_1. L_1 is used to find L_2, the collection of frequent 2-item sets, which is used to find L_3, and so on, until no more frequent k-item sets can be found. Finding each L_k requires a full scan of the database. With this process we can construct a collection of frequent item sets.

2.11.1 The *A priori* Algorithm Implementation

Implementation of the algorithm is another important issue in the research area. Using a priori implementation pseudo code, all the frequent item sets are determined from a number of parameters in a database transaction. In this code, *D* is the collection of database transactions, min-sup denotes the minimum support threshold, *L* is the number of frequent item sets in transaction *D* and C_i can become a member of the frequent item set. The following pseudo code represents general implementation method for A priori algorithm (Agrawal *et al*, 1993).

```
L1 → find frequent 1-item sets(D)
        For k in (1. Lk ≠ Ø, k++ ):
        Ck+1 → candidates for frequent item set generated
from
        (Lk with min-sup)
        Lk → L1 × Lk-1

        For each transaction t in D:
        increment the count of all candidates in Ck+1   if it
occurs in t
        Lk+1 → candidates in Ck+1 with min-sup
return ∪kLk
```

Important details of a priori algorithm

The purpose of this algorithm is frequent item set generation. It has a sub process which has an important role in the whole algorithm. The process has two main steps: First, for each L_k, the table will join $L_{k-1} \times L_1$. Second, the algorithm prunes the candidates which are not frequent and inside L_{k-1}.

2.11.2 Rule Generation from the *A priori* Algorithm

After generating the frequent item set, we will create rules from those items that have the highest frequency in the database. The second part of the association rule algorithm consists of two steps:

1. First, generate all subsets of S, in which S is the frequent item set.

2. Then, let SS represent a nonempty subset of S. Consider the association rule R: $ss \Rightarrow (s - ss)$. Generate (and output) R if R fulfills the minimum confidence requirement. Do so for every subset *ss* of *s*. Note that for simplicity, a single-item consequent is often desired.

2.11.3 Measures of Interestingness

After generating association rules, a possibly large number of rules will be generated. In general, the interestingness of a rule relates to the difference between the support of the rule and the product of the support for the antecedent and the support for the consequent. If the antecedent and consequent are independent of one another, then the support for the rules should approximately equal the product of the support for the antecedent and the support for the consequent. If the antecedent and consequent are independent, then the rule is unlikely to be of interest no matter how high the confidence (Piatetsky-Shapiro, 1991).

To reduce the number of rules, 'Lift' and 'Leverage' are two metrics that are used in the research as studied in the literature review.

a) Lift

'Lift' is described as the most popular measure for interestingness of a rule and is formulated as:

$$Lift\ (A \Rightarrow C) = \frac{Confidence\ (A \Rightarrow C)}{Support\ (C)} \qquad (2.3)$$

This is the ratio of the frequency of the consequent in the transactions that contain the antecedent over the frequency of the consequent in the data as a whole. If a lift value is greater than 1 then the consequent is more frequent in transactions containing the antecedent than in transactions that do not (Ibid, 1991).

b) Leverage

Another concept for rule interestingness measurement is 'Leverage' which is defined as:

$$Leverage\ (A \Rightarrow C) = Support\ (A \Rightarrow C) - Support\ (A) \times Support(C) \qquad (2.4)$$

Rules with higher leverage are more interesting than others. Measures such as lift or leverage can be used to further constrain the set of associations discovered by setting a minimum value. In addition, these measures have been used after rule generation because we need an antecedent and consequent for calculating its support and confidence so we cannot use them during the frequent item set calculation process.

2.12 Spatial Data

Geospatial data makes use of the geographic location of features and boundaries on Earth, such as natural or constructed features. Spatial data is commonly stored as coordinates and topology, and is data that can be mapped. Spatial data is often accessed, manipulated or analyzed through Geographic Information Systems (Rigaux *et al*, 2001).

Spatial data takes the form of Vector or Raster data. Vector data represents features through point, line and polygon data types, allowing the user to apply many relationships and geometrical concepts between them. On the other hand, raster data are in the form of matrix or array of data based on a pixel in such a way that each pixel has a value.

In general, both are used in spatial analysis but with different characteristics: vectors are good for spatial analysis of roads, areas, buildings etc., but raster data are good in calculations of and with neighbour pixels. In this research, we work with vector data to represent the topological concepts in the research project output and be capable of working with attribute data.

2.13 Spatial Data Mining

Nowadays, Spatial Data Mining (SDM) is a well-identified domain of data mining. It can be described as the discovery of interesting, implicit and previously unknown knowledge from large spatial databases (Han *et al*, 2006).

2.13.1 Topological Relationships in GIS

In the spatial use of data sets, an important concern is the spatial relation between objects. There are many types of relationships mentioned in the literature such as 'disjoint', 'contains', 'inside', 'equal', 'meet', 'covers', 'covered by', 'overlap' (Keating *et al*, 1987). Spatial topological relationships have a basic role in spatial analysis. In this section we describe some of these concepts, which are used in this research project:

- **Contains / Inside:** These types of relationships happen when a spatial object is completely covering the other. These concepts are most understandable with two polygon objects. If one of them is completely located inside the other one, then the relationship is 'contains'. In this sense if we change the situation of two objects we will achieve 'inside' relationship.

- **Close to:** The advanced types of spatial relationships are derived from the basic concepts with some additions. The term 'close to' is a kind of disjoint relationship with a specific threshold.

Other types of relationships and analysis options within GIS between spatial objects include table joins, buffers and overlays. Table joins are features of relational databases. Joining tables enables combination of data from multiple sources in analysis. Overlay operations in a GIS enables analysis between various layers of information. It thus facilitates numerous business applications as a result of the ability to query multiple map layers. A buffer on the other hand is an area surrounding a point, line or area defined by a radius distance. Buffers enables analysis of trade area as well as gauging the competition along, say, a lengthy commercial strip of which we do not discuss the details.

2.13.2 Spatial Association Rule

A spatial association rule is a rule in the form of A ⇒ B, where A and B are a set of predicates, some of which are spatial (Koperski *et al*, 1995). This definition gives a general idea about spatial association rule but there exist other definitions, which give a complete and specific schema to the concept.

A spatial association rule is a rule in the form of:

$$P_1 \wedge P_2 \wedge \ldots \wedge P_m \Rightarrow Q_1 \wedge Q_2 \wedge \ldots \wedge Q_n .\ (s\%, c\%) \qquad (2.5)$$

Where at least one of the predicates P_m or Q_n is a spatial predicate, s% is the support of the rule and c% is the confidence of the rule (Koperski, 1999). These concepts were discussed in Section 2.7. In spatial databases, certain topological relationships hold at all times (Egenhofer, 1991). They can be viewed as spatial association rules with 100% confidence. For example, the containment relationship expressed in Section 2.11 is one such association rule:

$$Contains(X, Y) \wedge contains(Y, Z) \Rightarrow contains(X, Z) \qquad (2.6)$$

However, such rules are usually domain-independent and therefore don't have meaningful information about specific database contents. An interesting spatial association rule may not always hold for all the data but may disclose some important spatial or topological features in the database.

2.14 Data Envelopment Analysis

Conceptually, Data Envelopment Analysis (*DEA*) is used to evaluate the efficiency of a number of producers. Typical statistical approaches are characterized as central tendency approach and evaluate producers relative to an average producer.

In contrast, *DEA* compares each producer with only the 'best' producers. In the literature, there are other definitions of *DEA* such as 'Data envelopment analysis provides a means of calculating apparent efficiency levels within a group of organizations. The efficiency of an organization is calculated relative to the groups observed best practice'. In the DEA literature, a producer is usually referred to as a decision making unit or DMU.

DEA was first described by Charnes, Cooper and Rhodes (CCR), and they demonstrated how to change a fractional linear measure of efficiency into a linear programming model (Ramathan, 2003). *DEA* is a mathematical programming model applied to observation data, which provides a new method of obtaining empirical estimates of external relations, such as the production functions and/or efficient production possibility surfaces that are fundamental to modern economics.

The efficiency of each decision making unit is a function of the amount and number of inputs and outputs, and the number, type, and characteristics of decision making units. In this sense, at the end, a scalar is identified as the relative efficiency, representing the total situation of that unit (Divandari *et al* 2006, Hesseinzadeh *et al*, 2007).

2.14.1 Use of *DEA* in Spatial Data

DEA is normally used in financial or business organizations with many branches. In this method, the spatial factor is not involved in the mathematical models. On the other hand, each phenomenon by itself has a spatial factor which cannot be ignored. For efficiency, one must take into account the location parameter of the business branch, e.g., whether it is in a residential area or in a trade area. In addition, the spatial characteristics should be added to the inputs and outputs of the *DEA* model to better estimation of the efficiency measurement.

Needless to say, each organization applies a different strategy for its business branches, based on their location. For example, in a bank some branches are expected to act as a resource absorber, while some others will be active in providing loans. A branch in a residential area cannot give loans like one in a trade area. The concept is very simple but it is not yet modeled in the scientific domain. An important issue of this research is to find a proper combination model of both the mathematical and spatial issues, in the spatial rules association method.

One of the important issues that we deal with in the research project is the combination of spatial parameters beside the financial factors, to increase the accuracy of the efficiency measure. That means, a high weight will be given to spatial parameters inside the model, and also in deriving the spatial association rules to improve the approximation.

2.14.2 Support and Confidence using *DEA*

Beside the usual methods for measuring support and confidence of derived rules, there is a method called 'ranking discovered rules from data mining with multiple criteria by data envelopment analysis' (Chen, 2006). The general idea is that in association rules regardless of spatial or non-spatial point of view, many useful and useless rules are generated and by using a proper *DEA* model, all candidates (derived association rules) are ranked using the efficiency concept in decreasing order. The top N candidates are selected. The evaluation of this method with the common support and confidence shows a better result for the *DEA*-based method (Ibid, 2006). Although this method shows better results, due to the fact that it needs many additional processes in data gathering such as preparing a questionnaire, this method has not been used in this research project.

2.14.3 Related Topics and Other Disciplines

In recent years, the use of spatial data analysis in GIS has become very popular. Various dimensions in the spatial data and in addition, huge amounts of attributes in these data, allow scientists to generate methods and algorithms in special branches. Spatial economics is concerned with the allocation of resources over space and the location of economic activity. In this branch of science, location analysis focuses mostly on one economic question, namely, location choice. This is only one decision among a large number of economic decisions (Anselin, 1990). On the other hand, a variety of parameters in spatial data such as economical and social exist in spatial economics. Mathematical and statistical methods help to analyze spatial data while economical theories are combined with them (Anselin *et al*, 1992).

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Study Area

The study area selected for this research project is the city of Nairobi, capital of Kenya. The geographical location of Nairobi is about 1° 15′ South and 36° 45′ East. According to the last census data (1999), the population of Nairobi is approximately 2 million. The projected population as at 2009 is approximately 3 million. Nairobi has 49 Locations of which 31 locations cover the study area.

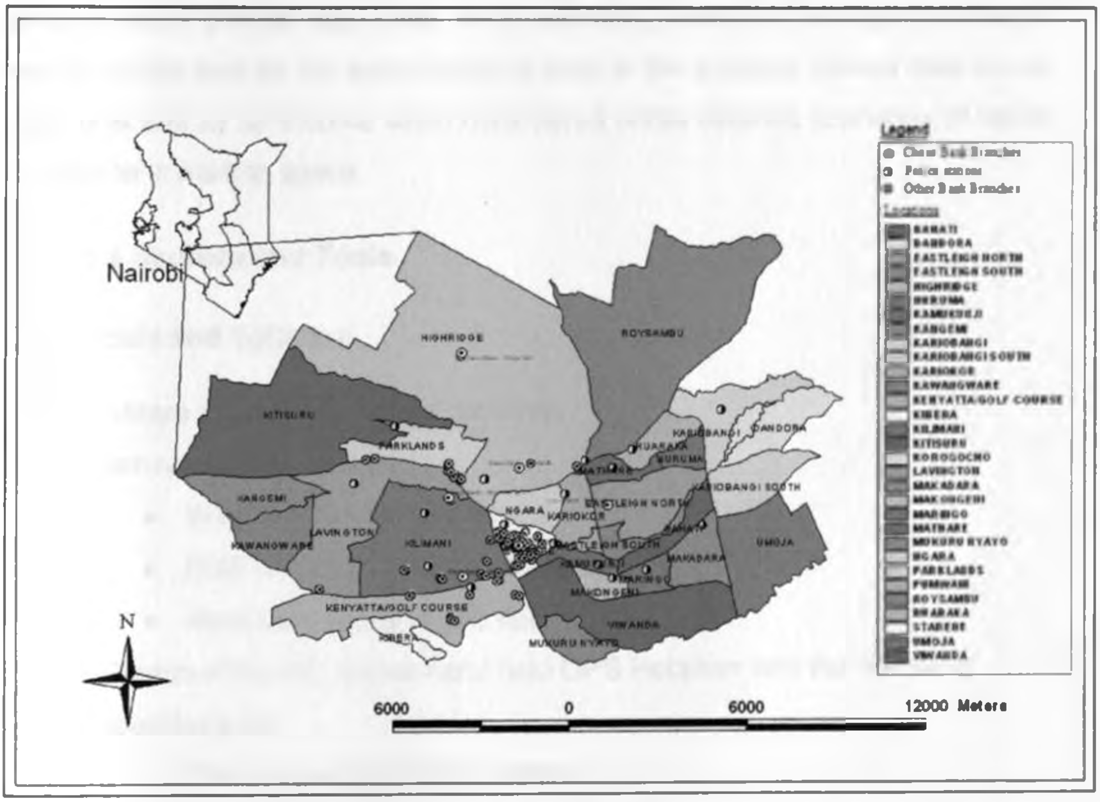


Figure 3.1: Study Area (Source – Survey of Kenya, 1999)

As mentioned in the first chapter, an application of this research model is in the banking industry. The study aimed to study and find relationships between bank branches as a case study by evaluating specific econometric data in relation to other local spatial parameters so as to generate association rules to guide expansion planning.

In general, when a bank opens a new branch in a certain area, the goal is to have an efficient branch, thus there is a direct connection between efficiency and site selection.

All parameters of the research have been chosen from a scientific background in site selection. Essentially, five categories are discussed in most research articles: population, competitors, access, land use, and income (Cliqquet, 2007). The structure of this research project in parameter perspective is extracted from these five classes. Due to the limitations of data gathering, in some cases, related parameters or proxies were used. As an example, instead of average income per region, rental data for the same region is used in the analysis. Rental data serves as a cost and as an income when considered under different scenarios of letting out space or leasing space.

3.2 Data Sources and Tools

3.2.1 Tools and Software

The hardware used in this project included:

1. Lenovo laptop
    - Windows XP, service pack 2
    - 2GB of Random Access Memory (RAM)
    - Hard Disk space of 130 GB
2. Garmin eTrex HC Series hand held GPS Receiver with the following specifications:
    - Positional Accuracy ±4 meters
    - Battery operated, Cable or PC/USB Adapter
    - Operating Temperatures -15º C to 70º C
    - Altitude – 17500 meters
    - Velocity – 0.1 meter/sec steady state

The software used in this project included:

1. ArcGIS 9.2
2. ArcView GIS 3.2

3. Efficiency Measurement System (Data Envelopment Analysis software)
4. *Weka* 3.6 Data Mining Software

3.2.2 Spatial Data

Data Required	Characteristics of Data	Source
Location data for existing bank outlets	Grid coordinates (Easting and Northing) referenced to WGS84 coordinate system and UTM projection	Data captured by use of a hand held GPS Receiver
Location data of existing police stations within the area of study	Grid coordinates (Easting and Northing) referenced to WGS84 coordinate system and UTM projection	Data captured by use of a hand held GPS Receiver
Population distribution data for Nairobi District	1999 census data with location as the smallest enumeration unit	Kenya National Bureau of Statistics
Road network data	Classified and unclassified roads	ILRI
Administrative map of Nairobi District	Map showing administrative units within Nairobi District. Location information was used in the analysis.	Survey of Kenya
Zoning map of Nairobi District	Zoning depicting land use information	Nairobi City Council
Rental data of the case study area	Floor area and cost per square foot for premises occupied by Chase Bank Branches.	Chase Bank (K) Ltd
Bank financial (econometric) data	Equity, transactions, customer deposits, revenue/income, expenses, number of accounts, customer details, profit and loss	Chase Bank (K) Ltd

3.3 Bank Data

3.3.1 Chase Bank

The main source data for banks in this research is Chase Bank (K) Ltd. It has 6 branches within the 31 locations that cover the study area. Using Global Positioning System (GPS) technology, all branch coordinates were measured with accuracy of ±4 meters. In addition to branch locations, non-spatial attributes of branches also have an important role in the research. In general, non-spatial attributes of the branches are used to calculate the efficiency. To do so, firstly, a relative comparison between branches called '*DEA* based efficiency' was generated. The result of this measurement is a normalized number that compares the effectiveness of a branch with the best branch. The concept was discussed in Section 2.14. This takes into account two additional more sensible measurements to clarify abstract efficiency concept. The total turnover/revenue and profit for each branch are supplementary selected information, for better explanation of the research model.

Obviously, there is no guarantee to have the same result with different indicators but as far as there is no absolute efficiency defined for branches, auxiliary measurements were employed to evaluate and validate the results.

Table 3.1 shows the relative branch efficiency measures benchmarked with Hurlingham Branch using Efficiency Measurement System, a *DEA*-Based Software.

It indicates the benchmarked efficiency scores derived from input parameters: [Assets, Equity and Employees] as well as output parameters: [Revenue and Profit]. DMU – Decision Making Unit represents the branch on which the relative efficiency measure is calculated. The results obtained were used in the subsequent comparative analysis and validation of the prediction model.

Table 3.1: Relative Efficiency Measures of Chase Bank Branches.

	DMU	Score	Benchmarks	Assets {Input}	Equity {Input}	Employees {Input}	Revenue {Output}	Profit {Outp
1	City Centre	405.83%	0					
2	Hurlingham	170.60%	6					
3	Parklands	38.63%	2 (0.12)	51931.30	0.00	2.52	0.00	1346
4	Eastleigh	83.27%	2 (0.18)	0.00	5441.78	12.53	0.00	1689
5	Riverside	30.40%	2 (0.02)	0.00	6018.69	1.81	0.00	191
6	Village Market	72.34%	2 (0.66)	0.00	6198.02	2.37	0.00	61
7	Mombasa	89.98%	2 (0.34)	99922.90	0.00	18.64	0.00	2016
8	Consolidated	97.96%	2 (0.57)	0.03	283621	95.47	0.00	22416

Chase Bank (K) Ltd has a market share of less than 1% in comparison with the industry totals. This information and data as sourced from Chase Bank Risk Analysis Report (2009) is given in Table 3.2 and Figure 3.2.

Table 3.2: Chase Bank Market Share.

Chase Bank Market Share – 2008			
Particulars	**Industry Banks**	**Chase Bank**	**% Share**
	Kes. Million	**Kes. Million**	**Percentage**
Balance Sheet	1,157,812	10,300	0.89%
Pre-Tax Profits	42,954	247	0.58%
Customer Deposits	849,480	7,147	0.84%
Loans And Advances	611,502	5,139	0.84%

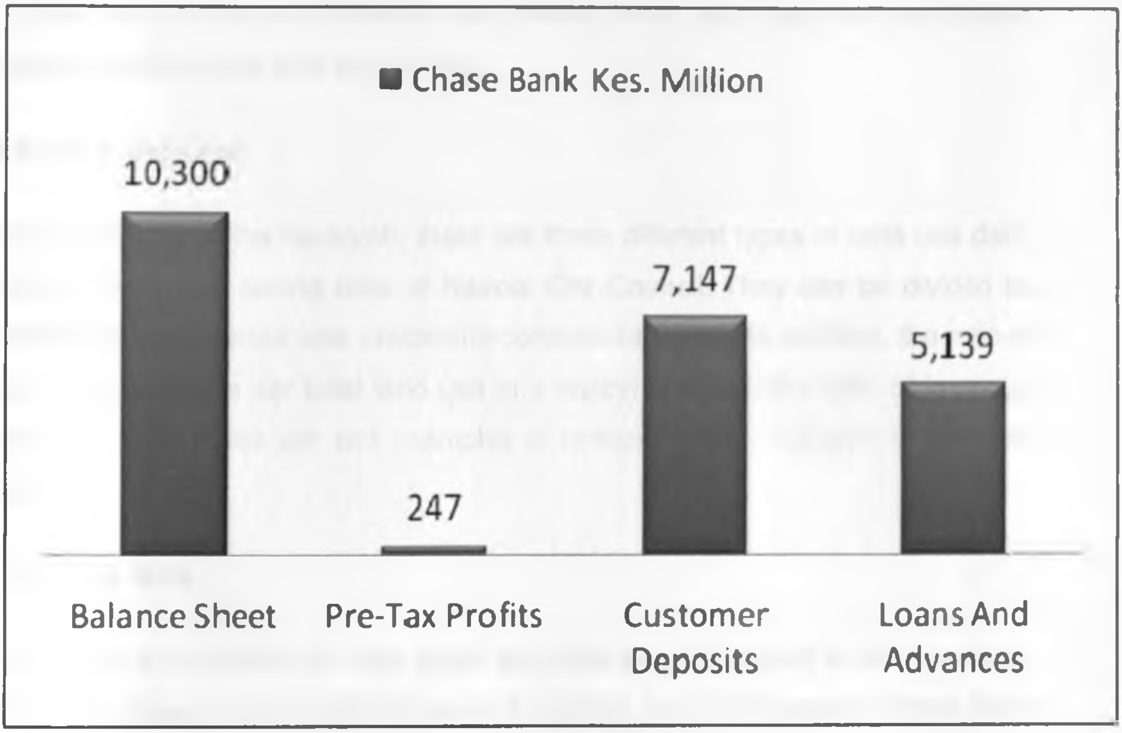


Figure 3.2: Chase Bank Market Share

3.3.2 Competitors

Two banks were selected as competitors for this category. Both were selected from the picked banks, which more or less have the same number of branches. The banks picked as competitors are Prime Bank and the ABC Bank. The parameter used in this category was competitor location. Non-spatial data for the competitors was avoided, because of some limitations. In other words, if the efficiency measure or the total turnover for the competitors were available, the result might be better.

3.4 Population Data

Various types of census data used in this research were sourced from the Kenya National Bureau of Statistics. Different attributes in census data consist of total population, poverty density, unemployed and literate people in each location. Furthermore, additional attributes were derived from main data such as ratio of discussed values per total population.

3.5 Land Use Data

For each scale in the research, there are three different types of land use data, derived from land zoning data of Nairobi City Council. They can be divided as residential, commercial and residential-commercial areas. In addition, the ratio of each land use size per total land use in a region and also the ratio of land use area per region area are two examples of complementary subjects in land use data.

3.6 Trade Area

In addition to the land use data, trade area has also been used in this research. The trade area was extracted by use of a multiple ring buffer around Chase Bank branches and clipping it with the population layer. The resultant layer as shown in Figure 3.3 has been used to depict the *Trade Area* that is covered by the bank and was used alongside the competition data to carry out spatial analysis.

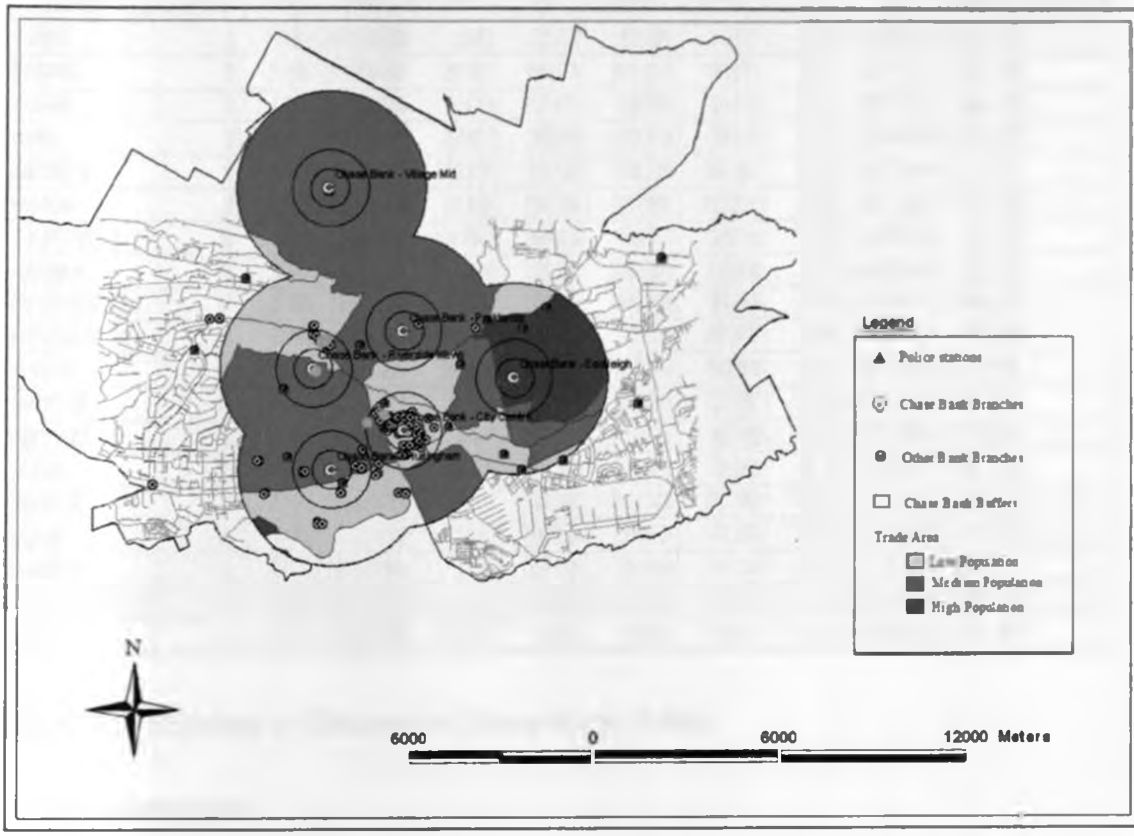


Figure 3.3: Chase Bank Trade Area with Population Information

Table 3.3 shows a sample statistical analysis done in Arc View 3.2 on distance to bank outlets based on trade area clipped shape with population range in the identified zones. The locations with minimum distance of 0.0000 meters as highlighted indicate the zones where the bank is represented in form of a branch outlet.

Stats of Distance to Bank_outlets.shp Within Zones of Pop_trdarea.shp

Location	Zone-code	Count	Area	Min	Max	Range	Mean	Std	Sum	[illegible]
STAREHE	1	2363914	2608573.7500	0.0000	829.7077	829.7077	176.9721	136.3189	418346880.00	8223 - 19361
KARIOKOR	2	2190376	2417075.0000	256.5612	1948.1735	1691.6123	1098.0753	373.9131	2405197824.0	27743 - 41752
MATHARE	3	576822	636521.7500	162.5014	1121.5714	959.0700	679.7202	231.0079	392077568.00	53006 - 80719
NGARA	5	2444372	2697359.0000	204.0627	1355.2418	1151.1791	858.4396	252.3754	2098345856.0	19362 - 27742
MAKONGENI	6	648329	715429.5625	1252.4187	2724.7385	1472.3198	2067.9697	321.0308	1340724864.0	8223 - 19361
VIWANDA	9	2980608	3289094.2500	134.4033	2352.3064	2217.9031	1109.2133	505.9541	3306129920.0	41753 - 53005
MUKURU NYAYO	10	71818	79251.0000	1098.5940	1649.4133	550.8193	1370.9110	131.2514	98456080.000	27743 - 41752
ROYSAMBU	12	567264	625974.5625	286.5699	1961.2371	1674.6672	1118.0894	379.0198	634251840.00	19362 - 27742
EASTLEIGH NORTH	13	574085	633501.5000	773.8258	2215.7874	1441.9615	1464.2964	359.6366	840630528.00	53006 - 80719
EASTLEIGH SOUTH	14	174979	193088.9219	1910.3785	2210.4490	300.0704	2051.8787	67.0845	359035680.00	41753 - 53005
PUMWANI	15	450735	497385.0625	691.4424	2066.5439	1375.1016	1542.9976	304.1681	695483008.00	8223 - 19361
KAMUKUNJI	17	1054027	1163116.3750	98.1056	2349.9902	2251.8848	1226.8387	618.0098	1293121152.0	8223 - 19361
PARKLANDS	18	3598110	3970506.2500	0.0000	1767.7797	1767.7797	826.1915	407.0263	2972728064.0	8223 - 19361
KITISURU	19	1596385	1761607.2500	463.3976	2440.3110	1976.9136	1728.5935	448.1806	2759500800.0	19362 - 27742
HIGHRIDGE	20	19558967	21583276.000	0.0000	2988.9485	2988.9485	1321.1884	675.1408	25841078272	27743 - 41752
KILIMANI	21	11947695	13184254.000	0.0000	1999.0704	1999.0704	764.8833	457.6543	9138592768.0	27743 - 41752
LAVINGTON	22	3112810	3434978.7500	0.0000	2097.0188	2097.0188	878.3244	482.4395	2734056960.0	8223 - 19361
KENYATTA/GOLF COUR	23	4450756	4911399.0000	0.0000	1152.6469	1152.6469	457.3217	231.5170	2035427584.0	19362 - 27742
KIBERA	24	132617	146342.5625	587.6701	1163.6746	576.0045	834.9647	131.4361	110730512.00	53006 - 80719

Table 3.3: Statistics of Distance to Chase Bank Outlets

3.7 Network Data

Network data is an important parameter for communication between business branches. In the location scales, it is classified in three different categories as 'highways', 'avenues' and street'. For each category, the total length inside the area is used as a measurement. In the branch scale, the access concept converts to the distance between an internal branch and the competitors and the distance to a police station as an urban facility.

Figure 3.4 shows the road network within the area of study with 100 meter multiple buffers along the classified roads which depict the connectivity of the selected bank branches. Chase Bank branch multiple ring buffer overlay has been included for ease of identification and interpretation.

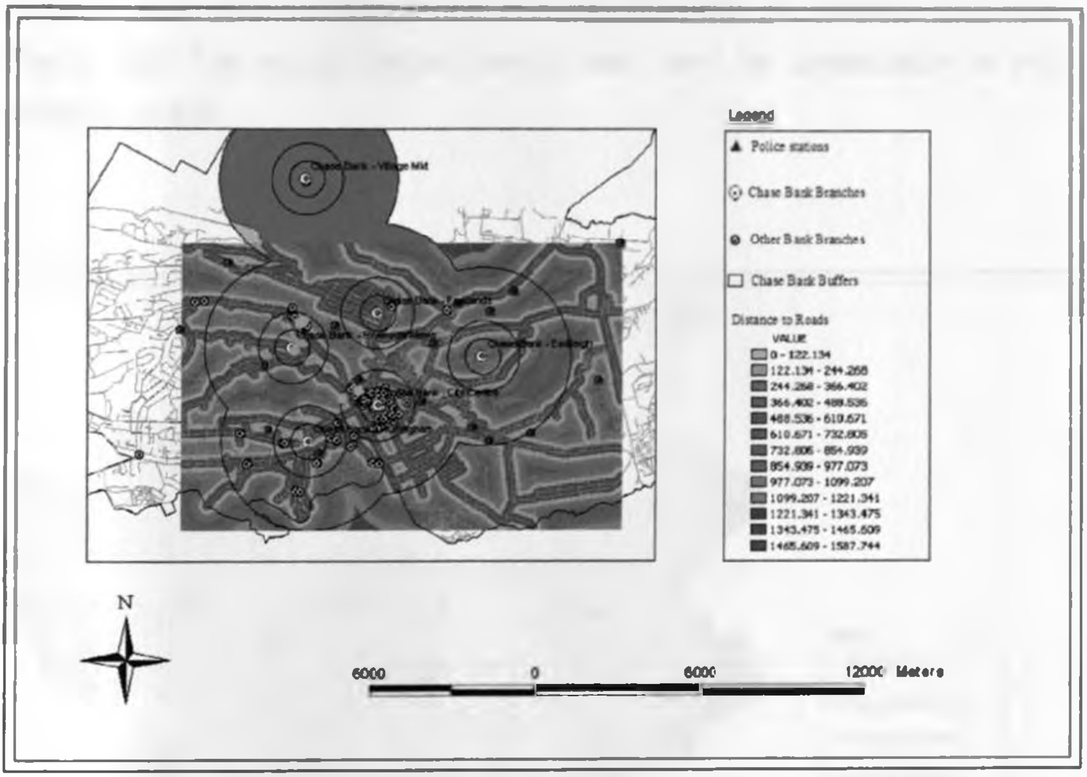


Figure 3.4: Road Network with 100 meter Buffers within the Study Area

3.7.1 Shortest Path In Network Data Using Buffer Rings

From the network parameter perspective, in the branch scale, additional network calculations for two point parameters have been used. The shortest path concept was used to find the minimum distance between bank branch points and urban facilities such as police stations. In this research, a buffer ring of 2500 meters has been used to determine the nearest police station to each branch. The calculated distance was then used for classification in three classes as 'high'- 2,500m, 'average'- 1000m and 'low'- 500m distance (see Figure 3.5).

In addition, the distance between Chase Bank (K) Ltd as an internal bank in comparison to the distance to the competitors has been used to determine the shortest path in the network. With a buffer of 2500 meters, minimum distance between them was measured, and also classified like the previous one (see Figure 3.6). The natural breaks method was used for classification in this category as well.

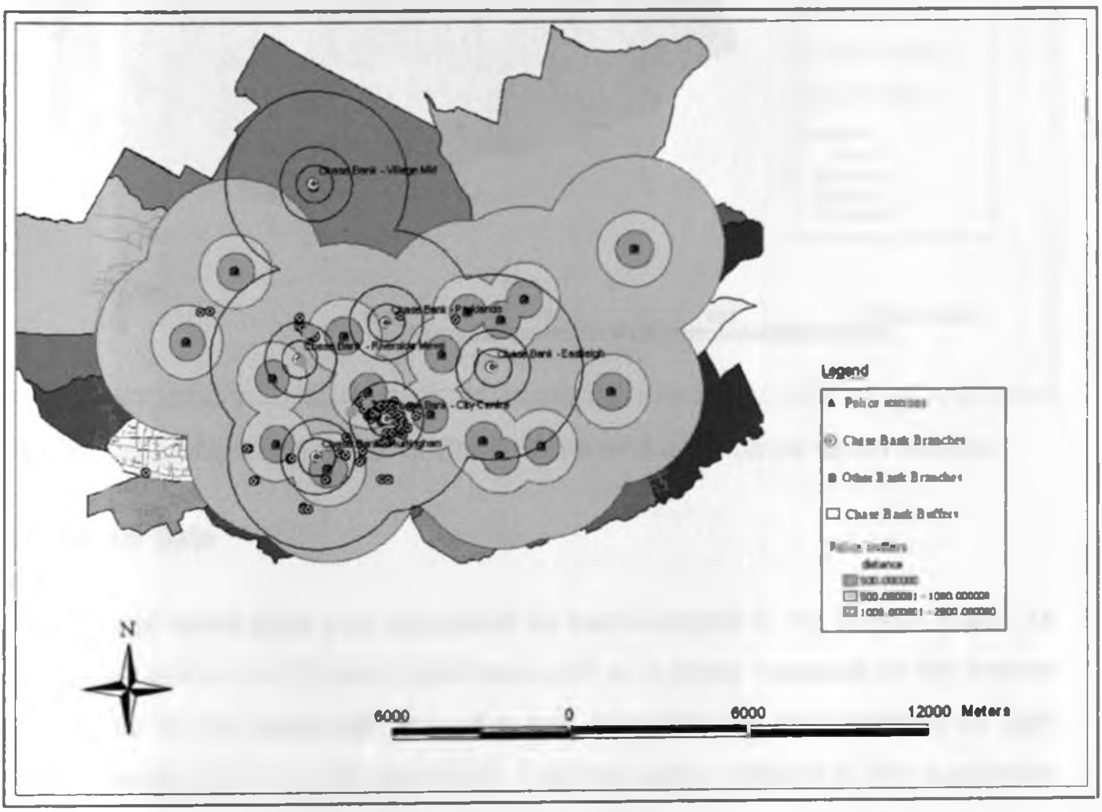


Figure 3.5: Ring buffer showing distance to police stations with Chase Bank overlay.

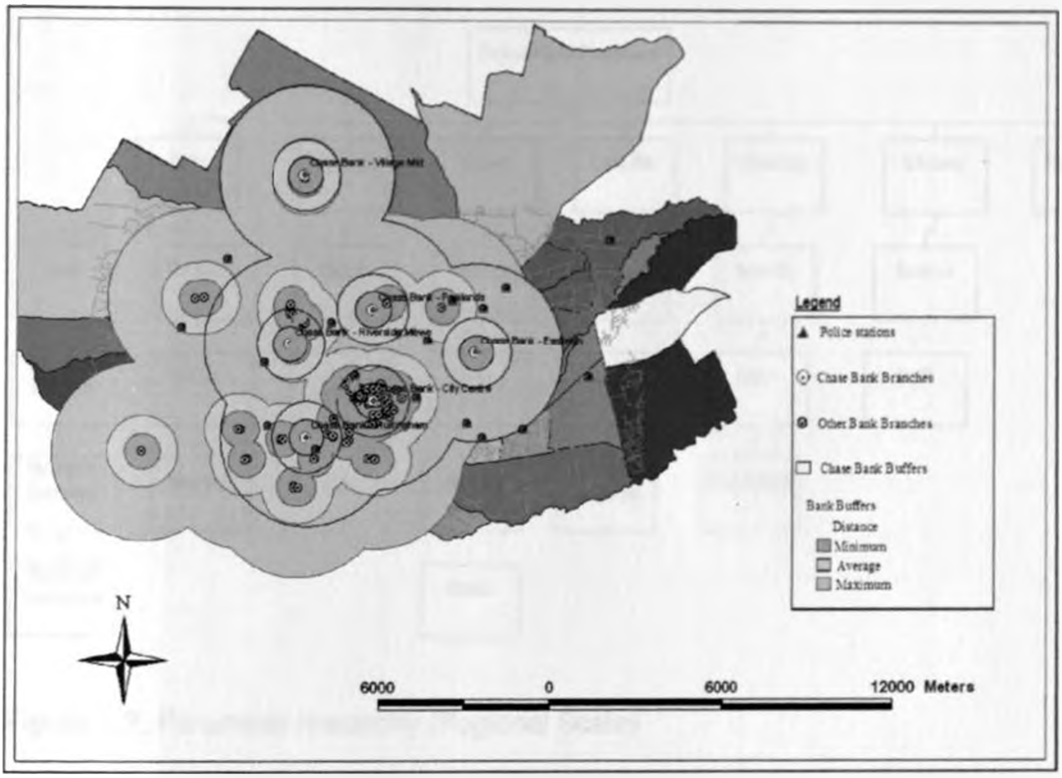


Figure 3.6: Multiple ring buffer of Chase Bank and other banks as an overlay.

3.8 Rental Data

An average rental price was calculated for each location in the branch scale. As discussed before, rental price data was used as a proxy measure of the income parameter. In this parameter, the value was classified into three classes as 'high price', 'average price', and 'low price'. The information relating to this parameter was then incorporated with the branch cost in order to determine both the relative and abstract efficiency measures for the branch business units.

3.9 Parameter Hierarchy

In this research, the aim was to analyze the association between parameters in regional and branch scale. Different parameters used in each mentioned scales are shown in a hierarchy (see Figure 3.7 and 3.8). The parameter hierarchy gives the general overview of the spatial factors used in the study.

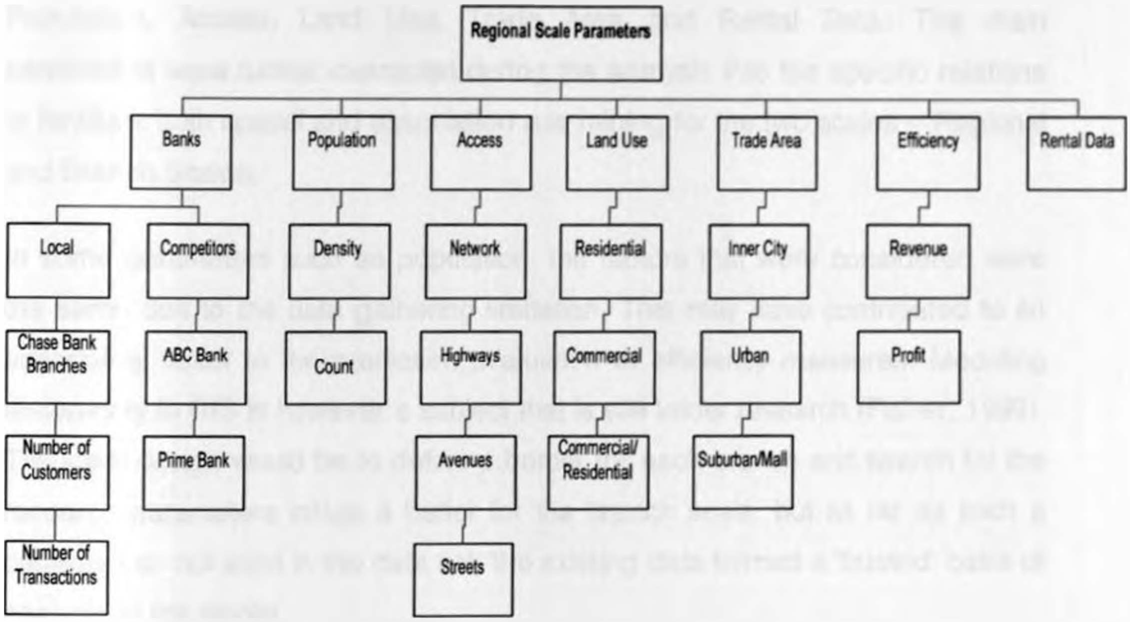


Figure 3.7: Parameter hierarchy (Regional Scale)

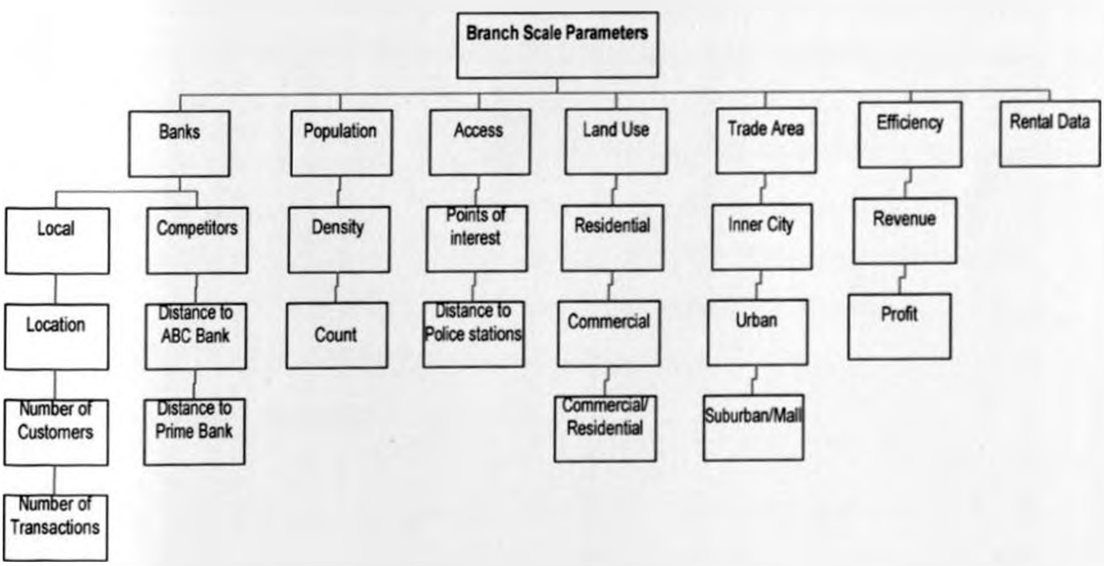


Figure 3.8: Parameter Hierarchy (Branch Scale)

The main parameters that were used in both scales are the same; Banks, Population, Access, Land Use, Trade Area and Rental Data. The main parameters were further cascaded during the analysis into the specific relations to facilitate both spatial and association rule mining for the two scales – Regional and Branch Scales.

In some parameters such as population, the factors that were considered were the same due to the data gathering limitation. This may have contributed to an uncertainty factor in the prediction evaluation of efficiency measures. Modeling uncertainty in GIS is however a subject that is still under research (Fisher, 1999). The ideal design would be to define a border for each branch and search for the research parameters inside a buffer for the branch scale, but as far as such a database do not exist in the data set; the existing data formed a 'trusted' basis of analysis of the model.

CHAPTER 4: RESULTS AND ANALYSIS

4.1 Overview

In this research project, the aim was to combine the non-spatial parameters with a set of different spatial parameters mentioned in the previous chapter, to find associations between the various parameters in order to predict efficiency at different spatial scales as well as finding the similarities and differences of derived spatial rules at both regional and branch scales.

In this chapter, a discussion of the results obtained is outlined. Besides, an *a priori* like algorithm of which the basic part was discussed in chapter 2, will be explained briefly as well as the two different scenarios related to the output of the algorithm.

4.2 Data Preprocessing

As discussed in the previous chapters, and regardless of output structure, a number of additional variables are derived from existing parameters. In this way, there is an early step in the method called data preprocessing, to prepare essential inputs for the main method. Extracted 'excel comma-delimited data' from the Bank's Central UNIX Database Server was converted to a *Weka.arff* data format to enable running of the *a priori* algorithm using *Weka* 3.6 Data Mining software. The converted data file was initially discretized using a non-supervised filter in the *Weka* 3.6 software environment. *Weka* is data mining software developed by the University of Waikato, New Zealand and is available for use for academic purposes free of charge.

Initially, the data file had 12 attributes which were later increased to 15 by an additional branch, turnover and efficiency attributes. Results generated by both files were compared, together with subsidiary Bank Risk Analysis data to validate the output and hence the loosely coupled model for predicting efficiency parameters for the potential sites in the study area.

4.2.1 Role of Efficiency Parameters in the Association Rule

At the early stages of the project, derivation of association rules in the whole study area was tried. However, because of the limited area coverage and the limited number of Chase Bank branches in the study area, the support and confidence of the rules generated were very low. The aim nonetheless is to find the potential sites in which the efficiency is 'high', 'average', or 'low'. The reason is that in determining the approximate areas where the above mentioned measure is nearly in the same range, it is assumed that efficiency measures will have a normal distribution in the said area and that the change in contiguous boundaries is smooth. To do so, a local Inverse Distance Weighted (IDW) polynomial interpolation method was used for this process.

Primary results shows that, in Inner City part of the study area, there is a spatial pattern for high efficiency, while the Northern part of the study area contains average efficient branches and there is a low efficient spatial pattern in the Central part of the study area (see Figure 4.1).

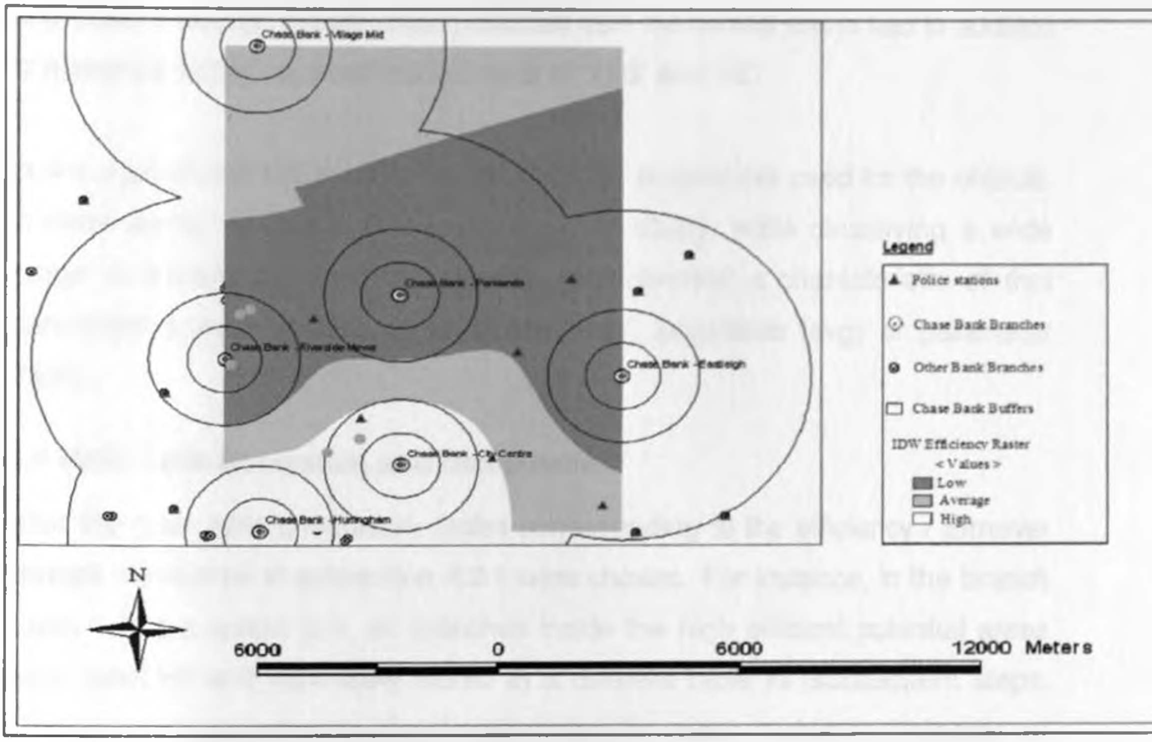


Figure 4.1: Spatial patterns of efficiency in the study area

4.2.2 Spatial Parameters Used in the Study

Spatial data used in this research originated from different social, economic, business and also infrastructure resources. At the city region and sub-region scales used in this research, there is a level of aggregation for some data. Point layers are combined in such a way that the total number of objects per region and sub-region are calculated in the process and also, polygon layers are merged at the sub-region level.

4.3 Different Types of Classification

For proper application of the *a priori* algorithm and also because of the wide range of data in the research project, all parameters were classified. Natural breaks method was used to classify the data for all the parameters. After data preprocessing step, all the parameters had a three way classification: 'High','Ave' and 'Low'. However, the attributes extracted from the central server had in addition to numerical factors, a classification mode of 'YES' and 'NO'.

In the *a priori* method, there is no classification preprocess used for the objects. In other words, all data is Boolean and in this study, while classifying a wide range of parameters in three classes, there existed a characteristic of that parameter in the main table as 'parameter (low)', 'parameter (avg)' or 'parameter (high)'.

4.4 Main Table Generation and Manipulation

After the main table generation, tuples corresponding to the efficiency / turnover classes mentioned in subsection 4.2.1 were chosen. For instance, in the branch scale, using a spatial join, all branches inside the high efficient potential areas were selected and separately stored in a different table for subsequent steps. The same procedure was also implemented for other scales and classes of efficiency measures. At the end of this stage different tables were generated with specifically those tuples that have highest probability for each efficiency classes. Then, to be able to use an array data structure of the *a priori* like algorithm implementation, the parameters were converted into an array data structure. This stage was done using a query, of which the pseudo SQL is as follows:

Select id, array[parameter1,parameter2 ,...., parameter n], efficiency real number into scale_array_efficiency
From scale_maintable_high/avg/low

As a result, the input for the *a priori* like algorithm implementation was obtained. Basic part of the *a priori* algorithm was discussed previously in Section 2.11. Figure 4.2 gives the flow chart used in association rule mining.

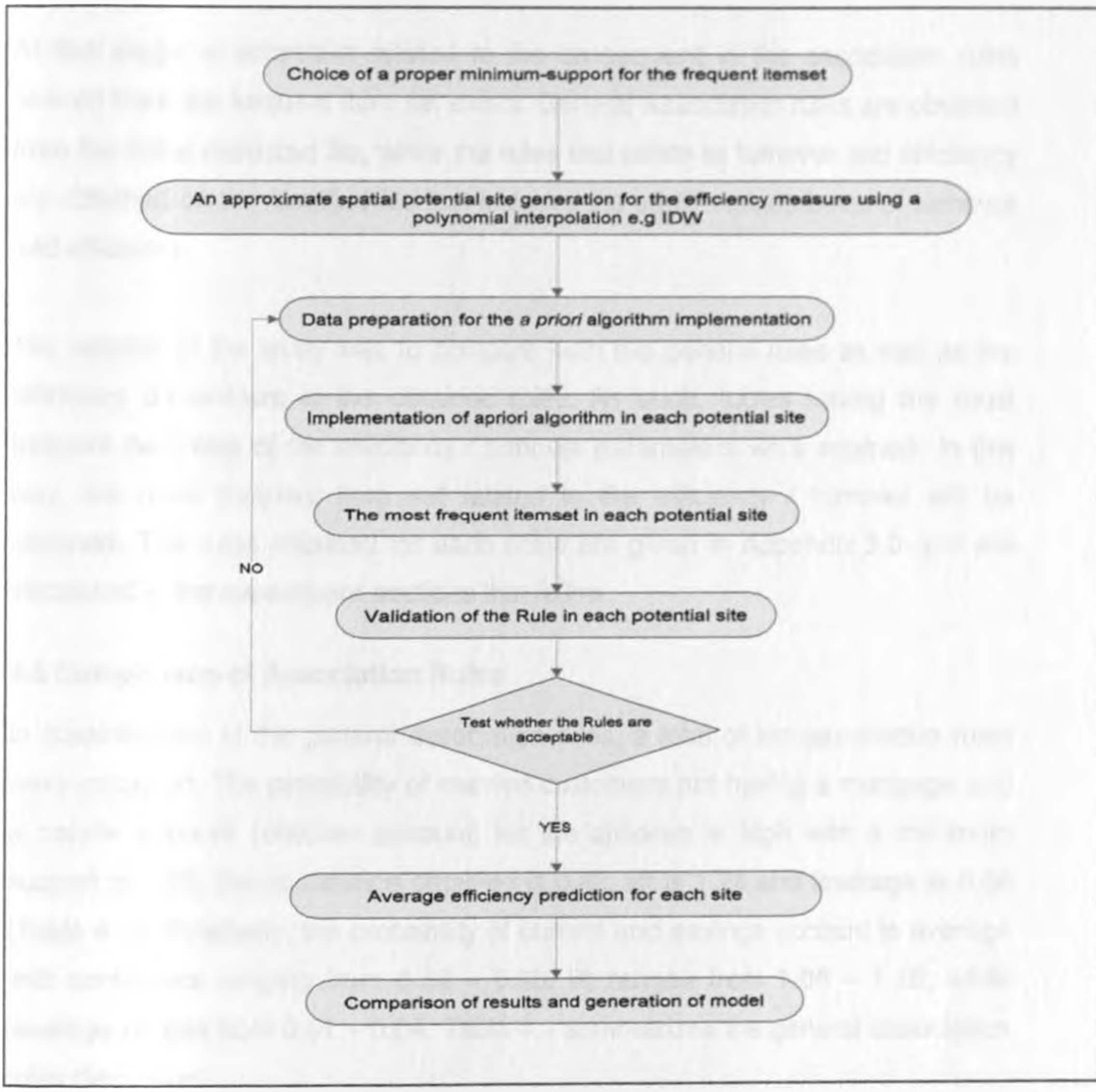


Figure 4.2: Flow Chart for the steps in association rule mining

4.5 Association Rule Generation

The next step was to find all possible association rules and calculate the support and confidence, lift and leverage for them. As discussed in subsection 2.11.2, for association rule generation there is a need to produce all subsets of the frequent item set. The outputs of the *a priori* like algorithm in each scale are candidates for association rule implementation.

At this stage, a constraint related to the consequent in the association rules derived from the frequent item set exists. General association rules are obtained from the initial extracted file, while the rules that relate to turnover and efficiency are obtained on the modified file that includes the additional attributes of turnover and efficiency.

The interest of the study was to compare both the general rules as well as the efficiency parameters in the obtained rules. As such, tuples having the most frequent item sets of the efficiency / turnover parameters were retained. In this way, the most frequent item set related to the efficiency / turnover will be obtained. The rules obtained for each scale are given in Appendix 3.0 and are discussed in the subsequent sections that follow.

4.6 Comparison of Association Rules

In consideration of the general association rules, a total of ten association rules were extracted. The probability of married customers not having a mortgage and a pepple account (*children account*) for the children is high with a minimum support of 0.25, the confidence obtained is 0.82, lift is 1.24 and leverage is 0.06 (Table 4.1). Relatively, the probability of current and savings account is average with confidence ranging from 0.52 – 0.80, lift ranges from 1.06 – 1.16, while leverage ranges from 0.01 – 0.04. Table 4.1 summarizes the general association rules discussed.

Table 4.1: General Association Rules Extracted

Association Rule	Confidence	Lift	Leverage	Probability
mortgage=NO pep=NO 209 ==> married=YES 171	0.82	1.24	0.06	high
save_act=YES pep=NO 235 ==> married=YES 175	0.74	1.13	0.03	average
married=YES mortgage=NO 261 ==> pep=NO 171	0.66	1.21	0.05	high
pep=NO 326 ==> married=YES 242	0.74	1.12	0.04	average
children='(-inf-0.3]' 263 ==> pep=NO 167	0.63	1.17	0.04	high
married=YES save_act=YES 277 ==> pep=NO 175	0.63	1.16	0.04	average
current_act=YES pep=NO 244 ==> married=YES 177	0.73	1.10	0.03	average
car=NO mortgage=NO 197 ==> current_act=YES 158	0.80	1.06	0.01	low
pep=NO 326 ==> married=YES mortgage=NO 171	0.52	1.21	0.05	high
married=YES pep=NO 242 ==> mortgage=NO 171	0.71	1.08	0.02	low

On the other hand, a total of 4 association rules were extracted for the turnover and efficiency parameters. Current accounts are associated with high rates of turnover and efficiencies in the rules generated. Comparing with the lift and leverage generated in the case of general association rules, it can be deduced that the probability on average is high for the rules. Table 4.2 summarizes the results obtained.

Table 4.2: Association Rules Based on Turnover and Efficiency

Association Rule	Confidence	Lift	Leverage	Probability
turnover=high 359 ==> efficiency=high 359	1.00	1.67	0.24	high
efficiency=high 359 ==> turnover=high 359	1.00	1.67	0.24	high
turnover=high 359 ==> current_act=YES efficiency=high 277	0.77	1.67	0.19	high
current_act=YES turnover=high 277 ==> efficiency=high 277	1.00	1.67	0.19	high
efficiency=high 359 ==> current_act=YES turnover=high 277	0.77	1.67	0.19	high

4.7 Auxiliary Data and Analysis

To complement and validate the results obtained from GIS analysis and Association Rule Mining using selected Data Mining Algorithms, an evaluation of econometric information (*non-spatial data*) extracted from the bank's Risk Analysis Survey Report for 2009 was carried out. The report covers both the analysis of the performance of the bank's branches as well as analysis and benchmarking of performance for the more than 40 banks in Kenya.

For the purpose of this study, information that relate to the bank's performance as well as the selected competitor banks have been extracted for comparison and validation of the results obtained using the scientific tools discussed in this research project. The key performance indicators (parameters) considered includes; efficiency, employee productivity, Return on Capital Employed (ROCE) as well as the profit and loss summary for the year 2008. These auxiliary data supports the performance predictive model in financial/economic terms and provides a linkage as well as a validation check of the association rules generated and GIS analysis carried in the previous part of this project report.

Table 4.3: Employee Productivity Ratio

Employee Productivity Ratio (Total Income/No of staff) -2008			
Name Of Institution:	Income Kes. Million	Number of Staff	Staff Cost Ratio
Prime Bank Limited	1,192	234	5.09
African Banking Corporation	725	171	4.24
Chase Bank Limited	764	220	3.47

Source - Chase Bank Risk Analysis Report (2009)

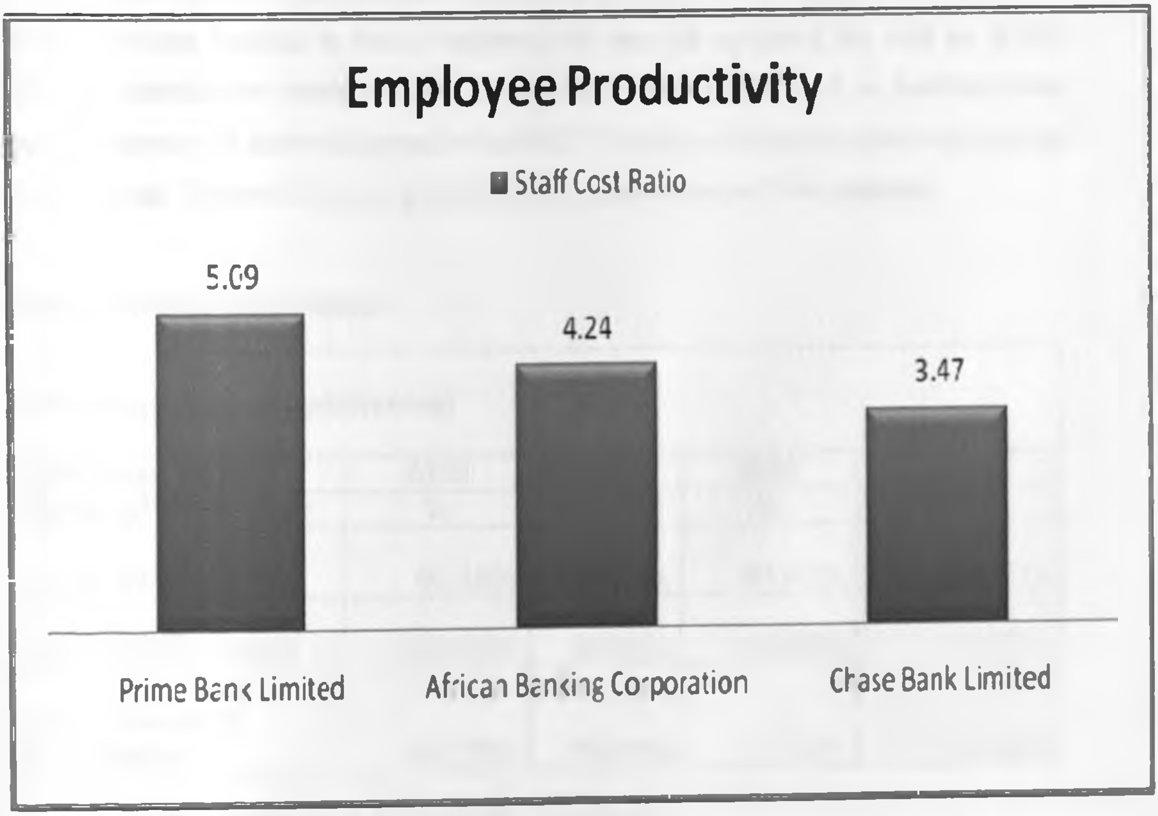


Figure 4.3: Employee Productivity Ratio

Table 4.3 gives a comparison of staff productivity ratios as a function of income for the three selected banks – Chase Bank Limited, Prime Bank and African Banking Corporation (the ABC Bank). Prime Bank and the ABC Bank were selected as competitors of Chase Bank as they are considered in the same tier/category of banks in terms of capital and asset strength.

When compared with the competitor banks selected, Chase Bank Limited continues to register a lower productivity index of 3.47 in terms of staff cost ratio compared with higher indices for both Prime Bank and ABC Bank (see Figure 4.3). This means that the bank spends more in terms of staff to generate income/turnover and hence gives an indication of an overall low efficiency in comparison with the competitors. Reasons that can be adduced to this relative low efficiency include a heavy reliance on manual systems as well as a thin branch distribution network in the trade /catchment trade area. A justified need for investment in technology and robust ICT Systems cannot be overemphasized in this case. Figure 4.3 gives a graphical representation of this analysis.

Table 4.4: Efficiency Ratios

Efficiency Ratio (Cost/Income)				
Efficiency Ratio	**2005**	**2006**	**2007**	**2008**
Name Of Institution:	**%**	**%**	**%**	**%**
Prime Bank Limited	60.38%	68.27%	61.41%	61.41%
Chase Bank Limited	62.51%	65.80%	63.52%	67.54%
African Banking Corporation	65.75%	72.55%	71.37%	69.38%

Source - Chase Bank Risk Analysis Report (2009)

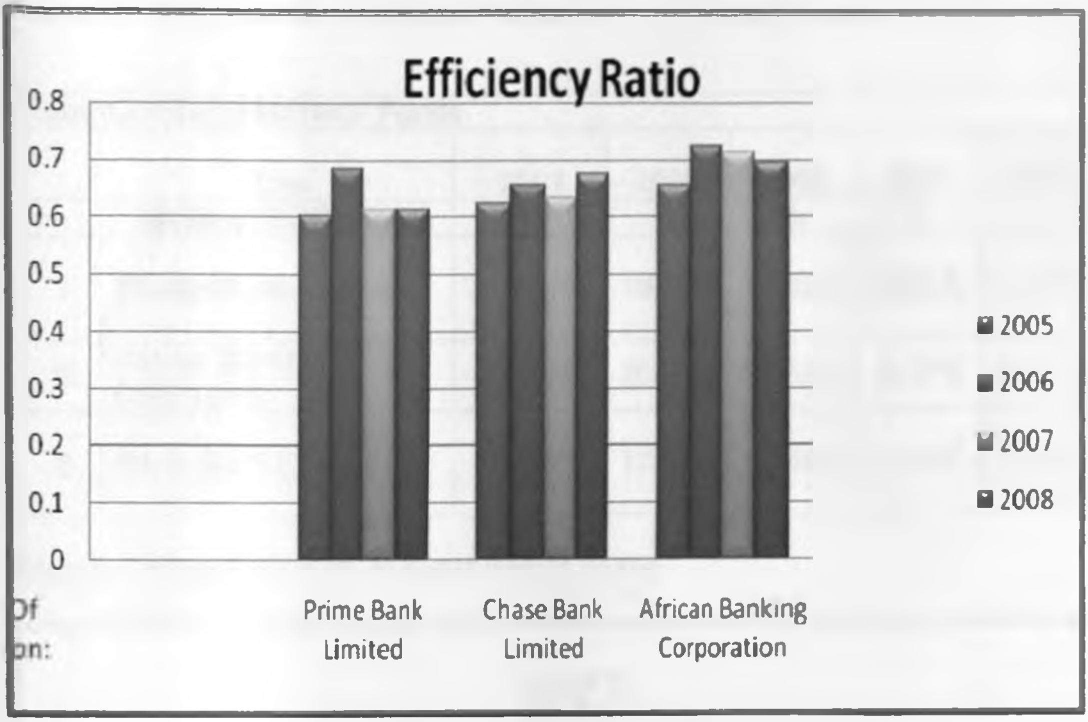


Figure 4.4: Efficiency Ratios – Cost vs Income

Building from the scenario displayed in Figure 4.3, cost-vs-income efficiency ratios as shown in Figure 4.4 depict a departure from an obvious correlation between staff productivity and efficiency measure. Both Prime Bank and the ABC Bank exhibit on average a declining/constant rate of efficiency measure. Chase Bank exhibits an increasing efficiency trend over the period under review. While it does not negate the need to invest in robust ICT Systems, it gives the indication that Chase Bank is well positioned in the niche market/trade area and can continue to exploit this opportunity by increasing the distribution network of branches and Automated Teller Machines (ATMs).

Table 4.5: Return on Shareholders' Funds

Return on Shareholders' Funds						
	Year	**2004**	**2005**	**2006**	**2007**	**2008**
	Name of Institution:	**%**	**%**	**%**	**%**	**%**
1	**Chase Bank Limited**	**-17.41%**	**10.88%**	**17.53%**	**25.81%**	**29.23%**
2	African Banking Corporation	23.52%	20.95%	20.68%	22.77%	22.93%
3	Prime Bank Limited	15.33%	17.32%	14.49%	16.44%	14.95%

Source - Chase Bank Risk Analysis Report (2009)

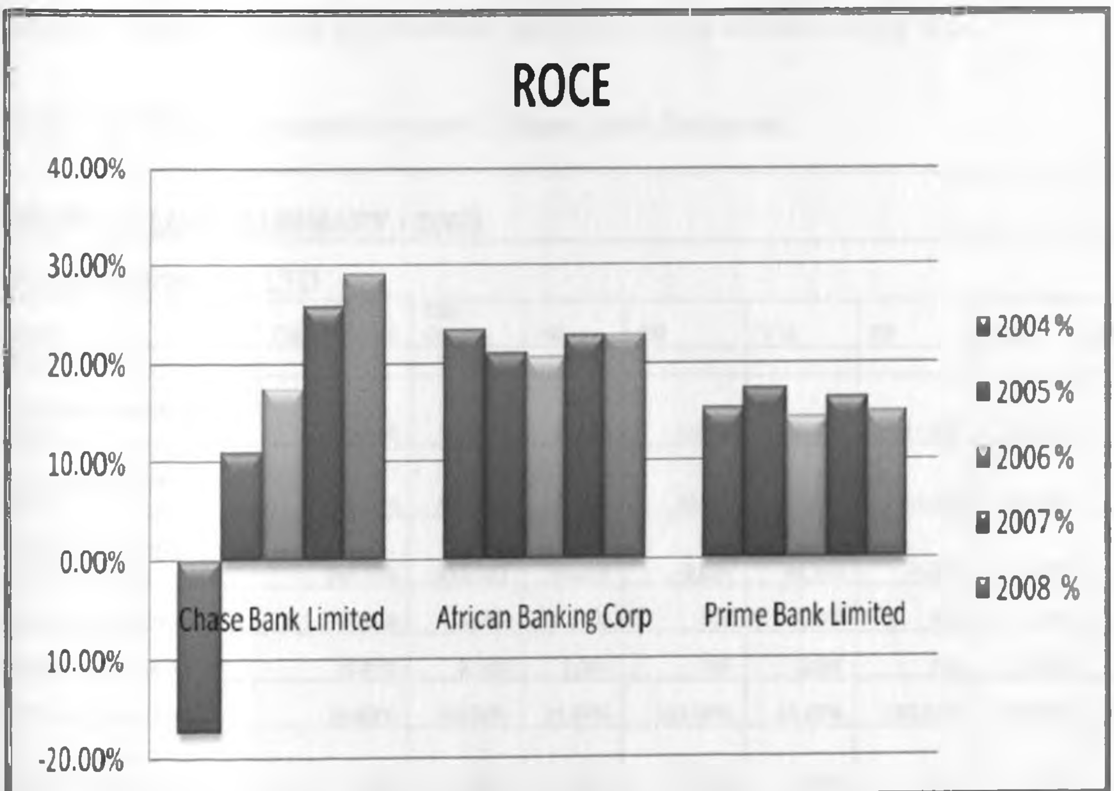


Figure 4.5: Return on Capital Employed

Efficiency on the basis of cost-vs-income presents a positive correlation with Return on Capital Employed (ROCE) as shown in Figure 4.5. Whereas, the ROCE for both Prime Bank and the ABC Bank depicts a constant trend, Chase Bank depicts an increasing ROCE which matches very well with the cost-vs-income efficiency measure.

Visually, as seen in Figure 3.3, the location of both Prime Bank and the ABC Bank branches are within 500m and 1000m from Chase Bank branches. Location parameters play a crucial role in banking competition and indicated in the efficiency raster map in Figure 4.1, higher rate of efficiency is recorded in the Inner City Region where banking competition is at a relatively higher level. This is indicated by the number of banks represented in this trade area, more so because of the business opportunities and commercial activities in the area.

Table 4.6: Profit and Loss Summary (*Chase Bank Branches*)

PROFIT & LOSS SUMMARY - 2008

CHASE BANK [K] LTD

Branch	Consolidated	City Centre	HB	EB	V M	PB	Msa	RM
TITLES								
Total Operating Income ('000)	749,140	542,410	89,103	15,762	58,571	11,031	30,244	2
Total Operating Expenses ('000)	424,212	327,089	19,297	19,397	14,222	15,403	26,334	2
OPERATING PROFIT / (LOSS) ('000)	324,928	215,321	69,806	-3,635	44,349	-4,372	3,910	
Number of Accounts	12,000	6,082	1,498	568	1,912	468	914	
Number of Transactions	16,032	8,126	2,001	759	2,554	625	1,221	
Efficiency (Cost/Income)	56.63%	60.30%	21.66%	123.06%	24.28%	139.63%	87.07%	122.
Relative Efficiency (Consolidated/Branch)	1.00	0.94	2.61	0.46	2.33	0.41	0.65	
Rank		3	1	6	2	7	4	

Data Source - Chase Bank Risk Analysis Report (2009)

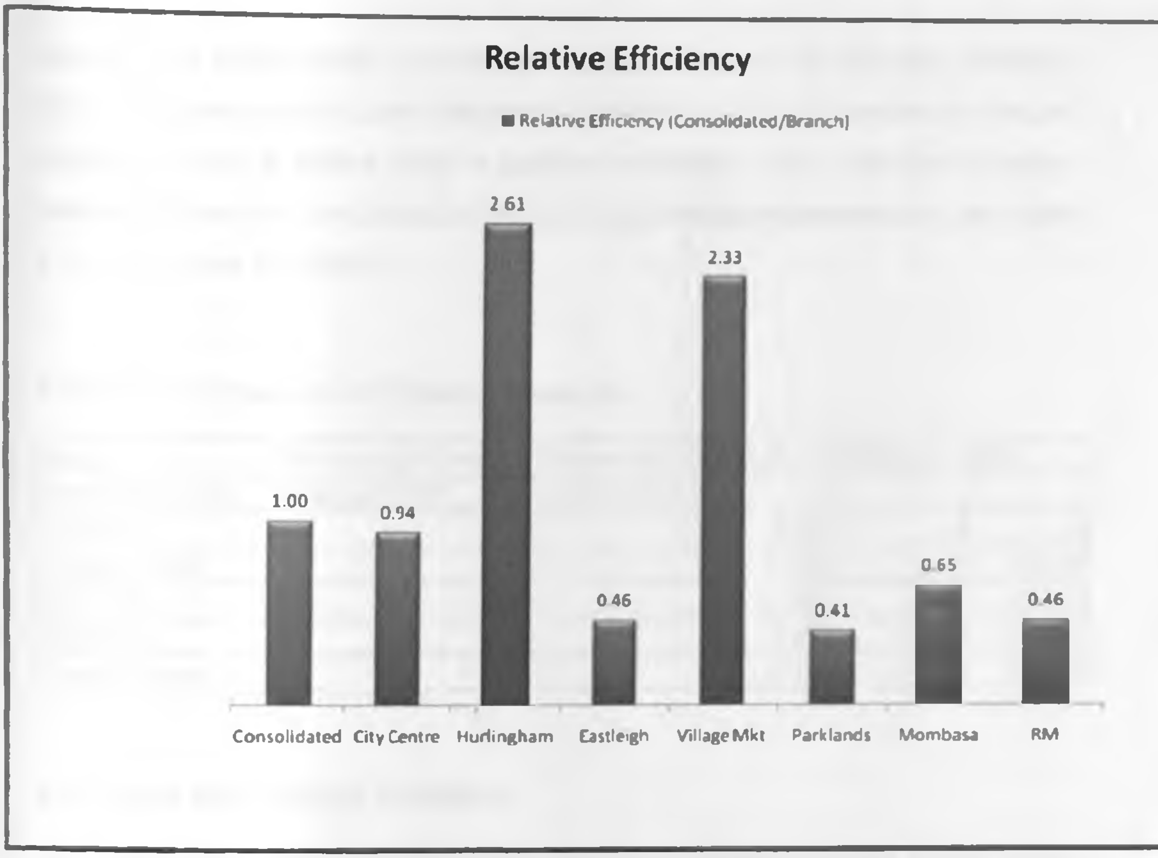


Figure 4.6: Profit and Loss Summary (Branches)

The relative efficiency measures as shown in Figure 4.6 gives an abstract comparison of the consolidated efficiency in relation to the contributory efficiency measure by the individual branches of Chase Bank Limited. The *DEA*-based efficiency measure tabulated in Table 3.1 gives a benchmarked efficiency measure with Hurlingham Branch, the best performing branch in terms of the both efficiency as well as the level of turnover and profitability.

In Comparison with the generated benchmarked efficiency measures on cost-vs-income (*with Hurlingham branch as a benchmark*), the other branches respectively have the following efficiency measures: City Centre – 0.36, Eastleigh – 0.18, Village Market – 0.89, Parklands – 0.16, Mombasa – 0.25 and Riverside Mews – 0.18. These efficiency measures compare well with the *DEA*-based counterparts even though the measurement parameters are different. However, the parameters used in both instances contribute to the bottom-line in terms of profitability and therefore have a positive correlation. The standard deviation calculated from the comparison of the efficiency measures/parameters (see table 4.7) in this case is 0.02217.

Table 4.7: Comparison of Efficiency Measures

Branch	DEA-based efficiency	Econometric Efficiency	Difference	RMS
(Benchmark - HB)	Revenue/Profit	Cost/Income		
Eastleigh	0.18	0.18	0	0
Village Market	0.66	0.89	-0.23	0.0529
Parklands	0.12	0.16	-0.04	0.0016
Mombasa	0.34	0.25	0.09	0.0081
Riverside Mews	0.02	0.18	-0.16	0.0256

4.8 Region Scale Result Limitation

As a result, the *a priori* algorithm could not find a frequent item set for the region scale due to the nature of this data set. That means it found more itemsets with three elements but none of them was able to generate a frequent item set with four items based on the min-support used for this scale. Besides, most of these tuples contain efficiency parameters with similar number of frequencies. Thus, for this data set, this scale was not good to generate frequent item set and the process was continued with the remaining scales.

4.9 Efficiency Prediction

Another strategy in this project is not only to find and generate different rules for existing parameters, but also to determine and predict the efficiency range for those sub regions / branches that include the most frequent item sets. This means, that after finding the most frequent item set, the average amount of efficiency measure in the branch scale is calculated, and set as the approximate range of efficiency prediction for any new branch, if such a case happens. This scenario is also valid for the turnover measure.

There is another scenario for the sub region or any other scale smaller than the branch scale. In scales smaller than branch scale, efficiency of the branches inside those sub regions can be measured or rely on the number of the '*high*', '*ave*' or '*low*' efficient branches inside those zones. The concept for the turnover measure is simpler than the *DEA*-based efficiency. As far as the amount of income can be aggregated and summarized in a single digit, the prediction of income is the summation of the most frequent item set elements.

4.10 Integrated Analysis Model

The aggregated representation of the model envisaged in this research project is represented in Figure 4.7. The representation gives a simplified representation of the main steps that are involved in generation of the prediction model that can aid in strategic expansion planning. The model is however not automated, but rather uses a combination of various software and analysis methods in a loosely coupled manner. Further research needs to be carried out to be able to come up with a tight-coupled/embedded model that can aid in expansion planning decision making process.

Based on the results obtained, an integrated analysis model would therefore seek to combine the results obtained from GIS analysis, Association Rules and Econometric analysis. This gives a three-pronged analysis method with all the elements loosely coupled as indicated above.

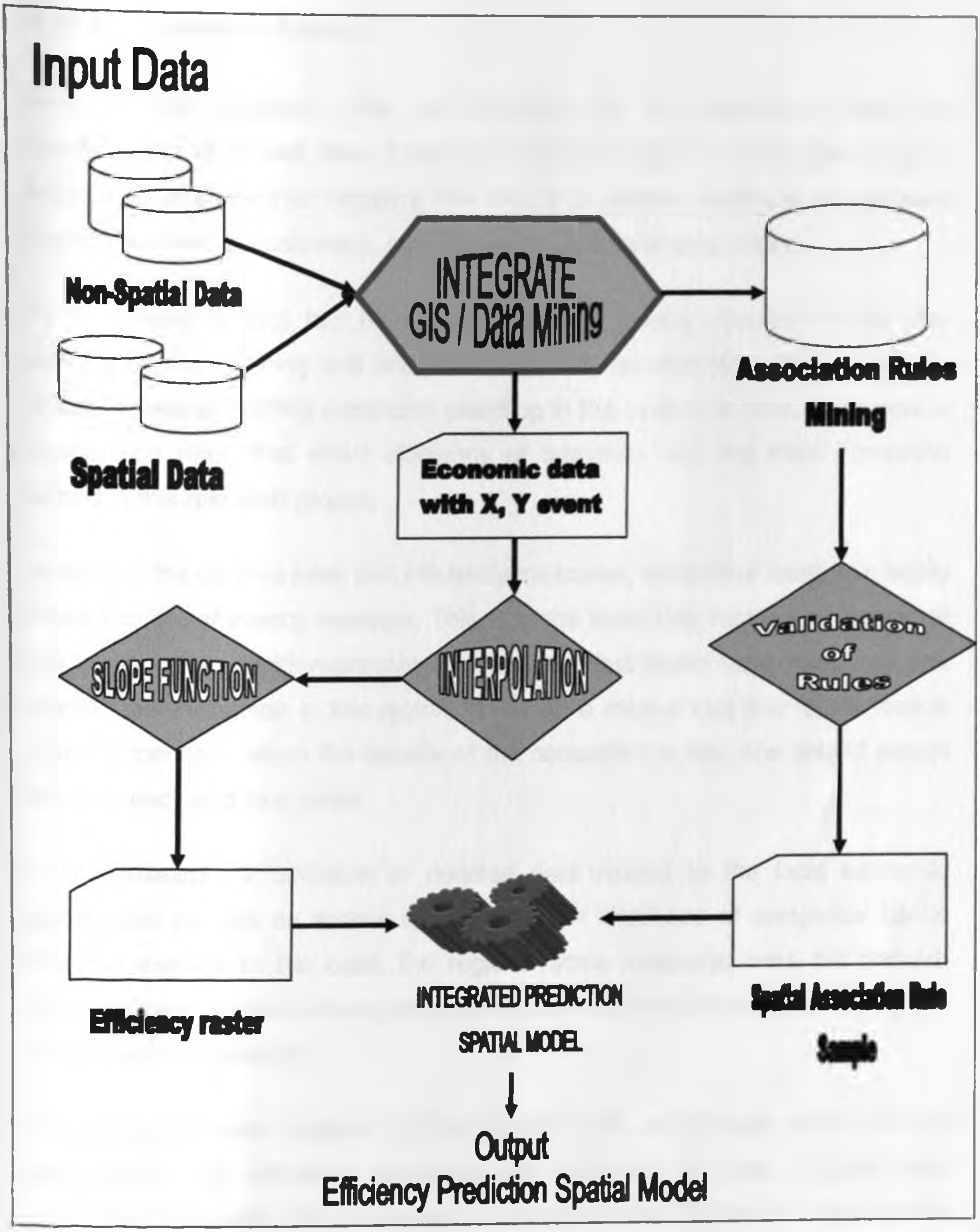


Figure 4.7: Proposed Integrated Analysis Model

4.11 Discussion of Results

GIS analysis functions offer an effective tool in generation, input and manipulation of spatial data. It can be used to integrate and analyse location factors/parameters that influence the choice of optimal locations for business outlets, location of customers, and demographic data among others.

In this research, GIS has been integrated in a loosely coupled manner with association rule mining and analysis of auxiliary econometric data to obtain a scientific way of guiding expansion planning in the case of a bank. Derivation of association rules that affect efficiency of branches was the main innovation aimed in this research project.

In most of the derived rules and efficiency measures, competitor location is highly related to the efficiency measure. This is in the Inner City region where most of the banks have branch representation. In fact, most banks have more than one branch representation in this region. It therefore means that due to the market sharing concepts, when the density of the competitor is low, one should expect low efficiency and vice versa.

In the research, a limitation of detailed data related to the local economic parameters as well as access to the customer database of competitor banks selected existed. In this case, the regional scale measures were not derived. However, density and location parameters were obtained and analyzed using the normal GIS functionality.

Comparing the rules obtained on the branch scale, a relatively higher lift and leverage for high efficiency parameter was obtained. Turnover measure also contributed rules with high measures of interestingness. This is as a result of the high relationship of turnover and profitability. These parameters contribute to the efficiency measure at the branch scale.

Using the method evaluated in this research, one can predict efficiency of branches in which the parameters are involved with the rules. In other words, due to the nature of the 'If then' rules, one can be able to predict efficiency measure when the antecedent happens. All the rules generated had leverage values of less than 1. In general however, lift values greater than 1 indicate that the consequent is more frequent in transactions containing the antecedent than in transactions that do not. This justifies that prediction of efficiency measures given the same nature of parameters in will always exhibit the general efficiency associated with the area of study. It thus validates the 'First Law of Geography' (Waldo Tobler, 1970) which has been widely used in spatial analysis which states that:

"Everything is related to everything else, but near things are more related than distant things."

This concept aptly explains that spatial relationships that have been used to model spatial trends of parameters in this research study. As described in section 4.10, the implementation of the loosely coupled analysis that incorporated GIS, Association Rule Mining as well as the validation of the efficiency parameters using econometric information can be automated to model a decision support system that can be used to guide strategic expansion planning. Such a model would involve an automated integration GIS analysis (spatial data), Econometric (non-spatial data) as well as Data Mining (association rule mining) to generate an Integrated Prediction Spatial Model.

This proposed automated integration model (see Figure 4.7) involves the merging of spatial and non-spatial data through the use of GIS and Data Mining to obtain Economic Data with (X, Y) event/components. The data can be interpolated through the use of polynomial interpolators like Kriging and IDW to obtain a slope function which represents an Efficiency Raster. On the other hand, Data Mining process yields association rules which can be in the form of either Spatial or Non-spatial Association Rule Samples. The data generated; Efficiency Raster and the Association Rule Samples are integrated through an algorithm to create a Spatial Prediction Model. This model can then be used to predict the candidate locations which can be evaluated to facilitate roll out of efficient branches/outlets.

Depending on the scale of operation, various parameters (see Figure 3.7 and 3.8) can be chosen based on the market dynamics of the candidate locations identified/generated from the prediction model. The best site is then selected from the candidate locations derived from the prediction model.

CHAPTER 5: CONCLUSIONS AND RECOMMENDATIONS

5.1 Conclusion

This research project aimed to find a prediction model for the efficiency parameters based on the spatial association rules. Using such a model, the managers of a business branch, are able to make better decisions for optimal allocation of new branches based on the derived rules.

Spatial association rule detection with a constraint is a method in which the outliers are not well-indicated. That means, with such a method, based on the *a priori* algorithm, one will miss the extreme cases and only spatial patterns that frequently happen will be obtained.

This method will predict the efficiency of areas which the spatial and non-spatial parameters are involved in the rules. If a responsible person suggests different areas for a new branch, the model will find the best one according to the highest validation concepts of the association rules such as lift and leverage. Experiments on this data set showed negative rules in certain classes. Such negative rules suggest areas which are not efficient. In comparison, the two measures in the derived rules show that the *DEA* based efficiency measure give a better result due to the number of relevant parameters in all classes.

The results obtained from the research give an alternative way to predict the average efficiency based on the most frequent item set which can help the managers to have a general idea about the new site. Efficiency measures derived will provide a scientific prediction of how a new branch will perform in the given site in relation to a selected benchmark – usually an existing branch with a relatively high efficiency measure.

An integration of spatial and non-spatial data in optimal site selection will offer objective means of evaluating candidate sites so as to help and guide managers make informed and strategic decisions on the best sites for business expansion planning and roll-out.

5.2 Recommendations

5.2.1 Recommendations for Practice

In view of the results and conclusions of this study the following recommendations were made:

- To be able to keep track of the changing IT environment, Banks need to put in place long term diagnostic process to identify the changes that are taking place and strategies that should be implemented.
- There should be adequate resource allocation to ICT strategic plans, commitment by all stakeholders and abandoning ideas that do not yield results.

5.2.2 Recommendation for Further Research

In this section, there are guidelines for any future works based on the experiences obtained from this research:

- Use this method for data sets with detailed characteristics of each building block. For branch scale there is need to calculate details of the spatial characteristics using a buffer for all points to have a better and complete result.

- A suggestion for the classification is to remove outliers or try to find a way to increase the number of elements in a high range class.

- Generate a complete automated process for the whole process including an interface to change the classification type and also parameter selection for the association rule. In addition, instead of classification method, there is need to find and test alternative methods for data sets which do not have a normal distribution and exhibit Poisson distribution. This will provide further insight and help validate the spatial association rules for discrete data that do not have a normal distribution.

- This study also recommends the use of an alternative method of spatial association rule based on parameter weight, where in the small scales, one is able to find a total weight of each parameter and due to that weight apply for the larger scales and implement the same rule by giving additional weights to some parameters.

- This research study was based on a snapshot of time for both the efficiency measurements and spatial parameters. This study strongly recommends addition of a time dimension to the research to come up with a temporal spatial association rule. In this case, one can also combine the temporal spatial association rule with a visualization method such as space-time-cube to detect the spatial temporal changes in the city.

REFERENCES AND BIBLIOGRAPHY

Agrawal, R. Imieliski, T. and Swami, A. (1993). Mining association rules between sets of items in large databases. In SIGMOD *'93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207-216, New York, NY, USA. ACM Press.

Anselin, L. (1990). What is special about spatial data? Alternative perspectives on spatial data analysis in spatial statistics: Past, present, and future. Ann Arbor, pages 63-77.

Anselin, L. and Getis, A. (1992). *Spatial statistical analysis and geographic information systems. Annals of Regional Science*, 26:19-33.

Armstrong, Marc P. (1992). *GIS and Group Decision-Making: Problems and Prospects*. GIS/LIS '92, vol. 1: 20-29.

Browning, J. (1990). *"Information Technology"*. American Economist Review.

Campbell, J. (2001). *Map Use and Analysis*. McGraw-Hill, fourth edition.

Chen, M. C. (2006). Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. *Expert Systems with Applications*.

Cliqquet, G. (2007). *Geomarketing*, 1st Edition, Brijpasi Art Press Ltd, Noida – India.

Craig,William J. and D. David Moyer (1991). *Progress on the Research Agenda*: URISA '90. URISA Journal, Volume 3, Number 1, Spring 1991, pp. 90-96.

Densham, P. J. (1991). *Spatial Decision Support Systems*. In: Maguire, D.J., M.F. Goodchild, and D.W. Rhind, eds. Geographical Information Systems: Principles and Applications, vol.1, London: Langman, 403-412.

Divandari, A. Jahanshahloo, G. R. and Hosseinzadeh Lotfi F. (2006). O. R. theory and its application, multi-component commercial bank branch progress and regress: An application of DEA. *International Mathematical Forum*, 1 (33):1635-1644.

Egenhofer, M. J. (1991). Reasoning about binary topological relations. In SSD *'91: Proceedings of the Second International Symposium on Advances in Spatial Databases*, pages 143-160, London, UK, Springer-Verlag.

Fisher, P. F. (1999). 'Models of Uncertainty in Spatial Data' in *Geographical Information System: Principles, Techniques, Management and Applications*, Longley P., Goodchild, M., Maguire D. and Rhind D. (eds), 1999, vol 1, New York, John Wiley & Sons pages 191 – 205.

Freeman & Perez (1988). *Structural Crises of Adjustment*, Business Cycles and Investor Behaviour.

Geoffrion, A.M. (1983). *Can OR/MS Evolve Fast Enough*? Interfaces 13: 10-25.

Goodchild M. F. and Zhou J. (2003). Finding Geographic Information: Collection level metadata. *Geoinformatica* 7(2), 95-112

Goodchild, M. F. and P. J. Densham (1990). Research Initiative Six. Spatial Decision Support Systems: Scientific Report for the Specialist Meeting. Technical Report 90-5. National Center for Geographic Information and Analysis.

Goodchild, M.F., R. Haining and S.Wise (1992). *Integrating GIS and Spatial Data Analysis: Problems and Possibilities*, International Journal of Geographic Information Systems 6(5): 407-423.

Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann, 2nd edition,

Hand, D. Mannila, H. and Smyth P. (2001). *Principles of Data Mining.* MIT Press, Cambridge.

Hosseinzadeh Lotfi, F., Navabakhs, M., Tehranian, A., Rostamy Malkhalifeh, M. and Shahverdi, R. (2007). Ranking bank branches with interval data: The application of DEA. *International Mathematical Forum*, 2 (9):429-440.

Jafrullah, M., Uppuluri, S., Rajopadhaye, N. and Srinatha Reddy, V. (2003). An integrated approach for banking GIS. Map India Conference, GISdevelopment.

Kargupta, H. Joshi, A. Sivakumar, K. and Yesha, Y. (2007). *Data Mining: Next generation challenges and future directions.* Prentice Hall of India, New Delhi.

Keating, T. Phillips, W. and Ingram, K. (1987). An integrated topologic database design for geographic information systems. *Photogrammetric Engineering and Remote Sensing,* 53(2):429-444.

Kenyaweb.com (2002). *Overview of the Kenyan Economy.* Available online at: www.kenvaweb.com/economy/overview/iondex.html

Koperski, K. (1999). *A progressive refinement approch to spatial data mining.* PhD thesis, Simon Fraser University.

Koperski, K. and Han, J. (1995). Discovery of spatial association rules in geographic information databases. In SSD '95: Proceedings of the 4th International Symposium on Advances in Spatial Databases, pages 47-66, London, UK, Springer-Verlag.

Korte, G. P. (2007). *The GIS Book: How to implement, manage and assess the value of Geographic Information System.* Akash Press, New Delhi.

Krygier J. and Wood, D. (2005). *Making Maps: A Visual Guide to Map Design for GIS.* The Guilford Press, new edition.

Larose, D. T. (2005). *Discovering Knowledge in Data: An Introduction to Data Mining.* John Wiley & Sons, Inc., Hoboken, New Jersey.

MacDonald, E. H. (2001). *GIS in banking: Evaluation of Canadian bank mergers,* Canadian Journal of Regional Science, XXIV(3):419-442.

Marakas, G. (2007). *Decision Support in the 21st Century.* Prentice Hall of India, New Delhi.

Miliotis, P., Dimopoulou, M. and Giannikos, I. (2002). A hierarchical location model for locating bank branches in a competitive environment. *International Transactions in operational Research,* 9(5):549-565.

National Research Council (1994). *Information technology in the service society.* National Academy Press, Washington.

NCGIA (1992). *A Research Agenda for Geographic Information and Analysis.* Technical Report 92-7. The National Center for Geographic Information and Analysis.

Ng, R. T. Lakshmanan, L.V.S. Han, J. and Pang, A. (1998). *Exploratory mining and pruning optimizations of constrained association rules.* In Proceedings of the 21st International Conference Management of data, pages 13-24.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In *Knowledge Discovery in Databases*, pages 229-248. AAAI/MIT Press.

Pick, J. B. (2005). *Geographic Information Systems in Business.* Idea Group Inc. USA

Ramanathran, R. (2001). *An Introduction to Data Envelopment Analysis.* SAGE Publications.

Rigaux, P. Scholl, M. and Voisard, A. (2001). *Spatial Databases: With application to GIS.* Morgan Kaufmann, New York, 2nd edition.

Rosenberg, N. (1994). *Exploring the Black Box.* Technology, Economics and History. Cambridge University Press, UK.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography 46:234–40.

Winograd, T. and F. Flores (1986). Understanding Computers and Cognition. Reading, Mass: Addison-Wesley.

Wit de, G.R. (1990). *The Character of Technological Change and Employment in Banking*: a Case study of Dutch Automated Clearing House (BGC).

APPENDICES

1.0 Rental Data

No	Branch Name	Shape	Number of tellers	Total Floor Area (sq. ft)	Rent per sq. ft (Kes)	Monthly Rent
1	City Centre	L-Shaped	4	2300	55/-	126,500/-
2	Parklands	Rectangular	3	1940	90/-	174,600/-
3	Hurlingham	Rectangular	4	3687	58/-	213,846/-
4	Eastleigh	L-Shaped	6	2659	119/-	316,421/-
5	Mombasa	Rectangular	6	5200	29/-	150,800/-
6	Riverside Mews	Square	4	2508	73/-	183,084/-
7	Village Market	Rectangular	2	1700	95/-	161,500/-
8	Thika	L-Shaped	8	5000	100/-	500,000/-

2.0 Sample Bank Data – Data Base Extract

Extract – Sample Bank Data

id	age	sex	trade_area	income	married	children	car	save_act	current_act	mortgage	pep
ID12101	48	FEMALE	INNER_CITY	17546	NO	1	NO	NO	NO	NO	YES
ID12103	51	FEMALE	INNER_CITY	16575.4	YES	0	YES	YES	YES	NO	NO
ID12112	52	FEMALE	INNER_CITY	26658.8	NO	0	YES	YES	YES	YES	NO
ID12116	38	FEMALE	INNER_CITY	22342.1	YES	0	YES	YES	YES	YES	NO
ID12119	62	FEMALE	INNER_CITY	26909.2	YES	0	NO	YES	NO	NO	YES
ID12121	61	MALE	INNER_CITY	57880.7	YES	2	NO	YES	NO	NO	YES
ID12123	54	MALE	INNER_CITY	38446.6	YES	0	NO	YES	YES	NO	NO
ID12125	22	MALE	INNER_CITY	12640.3	NO	2	YES	YES	YES	NO	NO
ID12126	56	MALE	INNER_CITY	41034	YES	0	YES	YES	YES	YES	NO
ID12127	45	MALE	INNER_CITY	20809.7	YES	0	NO	YES	YES	YES	NO
ID12129	39	FEMALE	INNER_CITY	29359.1	NO	3	YES	NO	YES	YES	NO
ID12134	33	FEMALE	INNER_CITY	29921.3	NO	3	YES	YES	NO	NO	NO
ID12136	27	FEMALE	INNER_CITY	19868	YES	2	NO	YES	YES	NO	NO
ID12142	47	FEMALE	INNER_CITY	26952.6	YES	0	YES	NO	YES	NO	NO
ID12145	20	MALE	INNER_CITY	13740	NO	2	YES	YES	YES	YES	NO
ID12146	64	MALE	INNER_CITY	52670.6	YES	2	NO	YES	YES	YES	YES

3.0 Data Mining Results

3.1 *A Priori* Run 1

```
=== Run information ===
Scheme:       weka.associations.Apriori -N 10 -T 3 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:     Bank_Data.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:    600
Attributes:   12
              Id,   age,   sex,   region,   income,   married,
children,   car,
              save_act,   current_act,   mortgage,   pep

=== Associator model (full training set) ===
Apriori - Run on 30/06/2009 at 12:59:34pm CAR - True
==================================
Minimum support: 0.25 (150 instances)
Minimum metric <conviction>: 1.1
Number of cycles performed: 15

Generated sets of large itemsets:
Size of set of large itemsets L(1): 16
Size of set of large itemsets L(2): 47
Size of set of large itemsets L(3): 16

Best rules found:

 1. mortgage=NO pep=NO 209 ==> married=YES 171    conf:(0.82) lift:(1.24) lev:(0.06) [33]
 2. save_act=YES pep=NO 235 ==> married=YES 175    conf:(0.74) lift:(1.13) lev:(0.03) [19]
 3. married=YES mortgage=NO 261 ==> pep=NO 171    conf:(0.66) lift:(1.21) lev:(0.05) [29]
 4. pep=NO 326 ==> married=YES 242    conf:(0.74) lift:(1.12) lev:(0.04) [26]
 5. children='(-inf-0.3]' 263 ==> pep=NO 167    conf:(0.63) lift:(1.17) lev:(0.04) [24]
 6. married=YES save_act=YES 277 ==> pep=NO 175    conf:(0.63) lift:(1.16) lev:(0.04) [24]
 7. current_act=YES pep=NO 244 ==> married=YES 177    conf:(0.73) lift:(1.1) lev:(0.03) [15]
 8. car=NO mortgage=NO 197 ==> current_act=YES 158    conf:(0.8) lift:(1.06) lev:(0.01) [8]
 9. pep=NO 326 ==> married=YES mortgage=NO 171    conf:(0.52) lift:(1.21) lev:(0.05) [29]
10. married=YES pep=NO 242 ==> mortgage=NO 171    conf:(0.71) lift:(1.08) lev:(0.02) [13]
```

3.2 *A Priori* Run 2

```
=== Run information ===
Scheme:     weka.associations.Apriori -N 5 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:   Main_Table.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:  600
Attributes: 15
        Id,        age,        sex,        trade_area,        income,        married,
children,
        Car,        save_act,        current_act,        mortgage,        pep,
branch,
        Turnover,        efficiency

=== Associator model (full training set) ===

Apriori - Run on 1/07/2009 at 15:12:37
===========================

Minimum support: 0.45 (270 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 11

Generated sets of large itemsets:
Size of set of large itemsets L(1): 12
Size of set of large itemsets L(2): 8
Size of set of large itemsets L(3): 1

Best rules found:

1. turnover=high 359 ==> efficiency=high 359    conf:(1) < lift:(1.67)> lev:(0.24) [144]
2. efficiency=high 359 ==> turnover=high 359    conf:(1) < lift:(1.67)> lev:(0.24) [144]
3. turnover=high 359 ==> current_act=YES efficiency=high 277    conf:(0.77) < lift:(1.67)>
lev:(0.19) [111]
4. current_act=YES turnover=high 277 ==> efficiency=high 277    conf:(1) < lift:(1.67)> lev:(0.19)
[111]
5. efficiency=high 359 ==> current_act=YES turnover=high 277    conf:(0.77) < lift:(1.67)>
lev:(0.19) [111]
```

3.3 *A Priori* Run 3

```
=== Run information ===

Scheme:     weka.associations.Apriori -N 10 -T 1 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:   Bank_Data.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:  600
Attributes: 12
        id          age          sex          region          income          married
children          car          save_act          current_act          mortgage          pep

=== Associator model (full training set) ===


Apriori - Run on 30/06/2009 at 12:56:04pm CAR - True
=================================

Minimum support: 0.25 (150 instances)
Minimum metric <lift>: 1.1
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 16

Best rules found:

 1. married=YES 396 ==> mortgage=NO pep=NO 171    conf:(0.43) < lift:(1.24)> lev:(0.06) [33]
conv:(1.14)
 2. mortgage=NO pep=NO 209 ==> married=YES 171    conf:(0.82) < lift:(1.24)> lev:(0.06) [33]
conv:(1.82)
 3. married=YES mortgage=NO 261 ==> pep=NO 171    conf:(0.66) < lift:(1.21)> lev:(0.05) [29]
conv:(1.31)
 4. pep=NO 326 ==> married=YES mortgage=NO 171    conf:(0.52) < lift:(1.21)> lev:(0.05) [29]
conv:(1.18)
 5. children='(-inf-0.3]' 263 ==> pep=NO 167    conf:(0.63) < lift:(1.17)> lev:(0.04) [24] conv:(1.24)
 6. pep=NO 326 ==> children='(-inf-0.3]' 167    conf:(0.51) < lift:(1.17)> lev:(0.04) [24] conv:(1.14)
 7. married=YES save_act=YES 277 ==> pep=NO 175    conf:(0.63) < lift:(1.16)> lev:(0.04) [24]
conv:(1.23)
 8. pep=NO 326 ==> married=YES save_act=YES 175    conf:(0.54) < lift:(1.16)> lev:(0.04) [24]
conv:(1.15)
 9. married=YES 396 ==> save_act=YES pep=NO 175    conf:(0.44) < lift:(1.13)> lev:(0.03) [19]
conv:(1.09)
10. save_act=YES pep=NO 235 ==> married=YES 175    conf:(0.74) < lift:(1.13)> lev:(0.03) [19]
conv:(1.31)
```

3.4 *A Priori* Run 4

```
=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 3 -C 1.1 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:    Bank_Data.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:   600
Attributes:  12
        id          age          sex          region          income          married
children
        car          save_act          current_act          mortgage          pep

=== Associator model (full training set) ===


Apriori - Run on 30/06/2009 at 12:59:34pm CAR - True
===================================

Minimum support: 0.25 (150 instances)
Minimum metric <conviction>: 1.1
Number of cycles performed: 15

Generated sets of large itemsets:

Size of set of large itemsets L(1): 16

Size of set of large itemsets L(2): 47

Size of set of large itemsets L(3): 16

Best rules found:

 1. mortgage=NO pep=NO 209 ==> married=YES 171    conf:(0.82) lift:(1.24) lev:(0.06) [33] <
conv:(1.82)>
 2. save_act=YES pep=NO 235 ==> married=YES 175    conf:(0.74) lift:(1.13) lev:(0.03) [19] <
conv:(1.31)>
 3. married=YES mortgage=NO 261 ==> pep=NO 171    conf:(0.66) lift:(1.21) lev:(0.05) [29] <
conv:(1.31)>
 4. pep=NO 326 ==> married=YES 242    conf:(0.74) lift:(1.12) lev:(0.04) [26] < conv:(1.3)>
 5. children='(-inf-0.3]' 263 ==> pep=NO 167    conf:(0.63) lift:(1.17) lev:(0.04) [24] < conv:(1.24)>
 6. married=YES save_act=YES 277 ==> pep=NO 175    conf:(0.63) lift:(1.16) lev:(0.04) [24] <
conv:(1.23)>
 7. current_act=YES pep=NO 244 ==> married=YES 177    conf:(0.73) lift:(1.1) lev:(0.03) [15] <
conv:(1.22)>
 8. car=NO mortgage=NO 197 ==> current_act=YES 158    conf:(0.8) lift:(1.06) lev:(0.01) [8] <
conv:(1.19)>
 9. pep=NO 326 ==> married=YES mortgage=NO 171    conf:(0.52) lift:(1.21) lev:(0.05) [29] <
conv:(1.18)>
10. married=YES pep=NO 242 ==> mortgage=NO 171    conf:(0.71) lift:(1.08) lev:(0.02) [13] <
conv:(1.17)>
```

3.5 *A Priori* Run 5

```
=== Run information ===

Scheme:      weka.associations.Apriori -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:    Bank_Data.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:   600
Attributes:  12
             id           age          sex          region          income          married
children
             car           save_act          current_act          mortgage          pep


=== Associator model (full training set) ===


Apriori - Run on 30/06/2009 at 13:02:46pm CAR - False
=================================

Minimum support: 0.1 (60 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 33

Size of set of large itemsets L(2): 161

Size of set of large itemsets L(3): 286

Size of set of large itemsets L(4): 171

Size of set of large itemsets L(5): 26

Best rules found:

 1. children='(-inf-0.3]' save_act=YES mortgage=NO pep=NO 74 ==> married=YES 73
conf:(0.99)
 2. sex=FEMALE children='(-inf-0.3]' mortgage=NO pep=NO 64 ==> married=YES 63
conf:(0.98)
 3. children='(-inf-0.3]' current_act=YES mortgage=NO pep=NO 82 ==> married=YES 80
conf:(0.98)
 4. children='(-inf-0.3]' mortgage=NO pep=NO 107 ==> married=YES 104    conf:(0.97)
 5. children='(-inf-0.3]' car=NO mortgage=NO pep=NO 62 ==> married=YES 60    conf:(0.97)
 6. married=YES children='(-inf-0.3]' save_act=YES current_act=YES 87 ==> pep=NO 80
conf:(0.92)
 7. married=YES children='(-inf-0.3]' save_act=YES mortgage=NO 80 ==> pep=NO 73
conf:(0.91)
 8. married=YES children='(-inf-0.3]' current_act=YES mortgage=NO 88 ==> pep=NO 80
conf:(0.91)
 9. sex=FEMALE married=YES children='(-inf-0.3]' mortgage=NO 70 ==> pep=NO 63    conf:(0.9)
```

3.6 A Priori Run 6

```
=== Run information ===

Scheme:     weka.associations.Apriori -I -N 10 -T 0 -C 0.9 -D 0.05 -U 1.0 -M 0.1 -S -1.0 -c -1
Relation:   Bank_Data.arff-weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-last
Instances:  600
Attributes: 12
        id          age          sex          region          income          married
children
        car          save_act          current_act          mortgage          pep


=== Associator model (full training set) ===


Apriori - Run on 30/06/2009 CAR at 14:02:11pm CAR - False (With item sets)
===========================================

Minimum support: 0.1 (60 instances)
Minimum metric <confidence>: 0.9
Number of cycles performed: 18

Generated sets of large itemsets:

Size of set of large itemsets L(1): 33

Large Itemsets L(1):
age='(-inf-22.9]' 60
age='(22.9-27.8]' 66
age='(32.7-37.6]' 62
age='(37.6-42.5]' 66
age='(42.5-47.4]' 71
age='(62.1-inf)' 68
sex=FEMALE 300
sex=MALE 300
region=INNER_CITY 269
region=TOWN 173
region=RURAL 96
region=SUBURBAN 62
income='(10825.799-16637.388]' 106
income='(16637.388-22448.977]' 110
income='(22448.977-28260.566]' 108
income='(28260.566-34072.155]' 76
income='(34072.155-39883.744]' 62
married=NO 204
married=YES 396
children='(-inf-0.3]' 263
children='(0.9-1.2]' 135
children='(1.8-2.1]' 134
children='(2.7-inf)' 68
car=NO 304
car=YES 296
save_act=NO 186
save_act=YES 414
current_act=NO 145
```

```
current_act=YES 455
mortgage=NO 391
mortgage=YES 209
pep=YES 274
pep=NO 326

Size of set of large itemsets L(2): 161

Large Itemsets L(2):
age='(62.1-inf)' save_act=YES 61
sex=FEMALE region=INNER_CITY 131
sex=FEMALE region=TOWN 92
sex=FEMALE married=NO 105
sex=FEMALE married=YES 195
sex=FEMALE children='(-inf-0.3]' 132
sex=FEMALE children='(0.9-1.2]' 66
sex=FEMALE children='(1.8-2.1]' 64
sex=FEMALE car=NO 153
sex=FEMALE car=YES 147
sex=FEMALE save_act=NO 94
sex=FEMALE save_act=YES 206
sex=FEMALE current_act=NO 70
sex=FEMALE current_act=YES 230
sex=FEMALE mortgage=NO 205
sex=FEMALE mortgage=YES 95
sex=FEMALE pep=YES 130
sex=FEMALE pep=NO 170
sex=MALE region=INNER_CITY 138
sex=MALE region=TOWN 81
sex=MALE married=NO 99
sex=MALE married=YES 201
sex=MALE children='(-inf-0.3]' 131
sex=MALE children='(0.9-1.2]' 69
sex=MALE children='(1.8-2.1]' 70
sex=MALE car=NO 151
sex=MALE car=YES 149
sex=MALE save_act=NO 92
sex=MALE save_act=YES 208
sex=MALE current_act=NO 75
sex=MALE current_act=YES 225
sex=MALE mortgage=NO 186
sex=MALE mortgage=YES 114
sex=MALE pep=YES 144
sex=MALE pep=NO 156
region=INNER_CITY married=NO 91
region=INNER_CITY married=YES 178
region=INNER_CITY children='(-inf-0.3]' 121
region=INNER_CITY children='(0.9-1.2]' 65
region=INNER_CITY car=NO 139
region=INNER_CITY car=YES 130
region=INNER_CITY save_act=NO 96
region=INNER_CITY save_act=YES 173
region=INNER_CITY current_act=NO 64
region=INNER_CITY current_act=YES 205
region=INNER_CITY mortgage=NO 175
region=INNER_CITY mortgage=YES 94
```

region=INNER_CITY pep=YES 123
region=INNER_CITY pep=NO 146
region=TOWN married=YES 115
region=TOWN children='(-inf-0.3]' 76
region=TOWN car=NO 82
region=TOWN car=YES 91
region=TOWN save_act=YES 128
region=TOWN current_act=YES 128
region=TOWN mortgage=NO 108
region=TOWN mortgage=YES 65
region=TOWN pep=YES 71
region=TOWN pep=NO 102
region=RURAL married=YES 61
region=RURAL save_act=YES 70
region=RURAL current_act=YES 72
region=RURAL mortgage=NO 68
income='(10825.799-16637.388]' married=YES 73
income='(10825.799-16637.388]' save_act=YES 71
income='(10825.799-16637.388]' current_act=YES 81
income='(10825.799-16637.388]' mortgage=NO 74
income='(10825.799-16637.388]' pep=NO 74
income='(16637.388-22448.977]' married=YES 77
income='(16637.388-22448.977]' car=NO 62
income='(16637.388-22448.977]' save_act=YES 65
income='(16637.388-22448.977]' current_act=YES 83
income='(16637.388-22448.977]' mortgage=NO 73
income='(22448.977-28260.566]' married=YES 70
income='(22448.977-28260.566]' save_act=YES 61
income='(22448.977-28260.566]' current_act=YES 75
income='(22448.977-28260.566]' mortgage=NO 65
income='(22448.977-28260.566]' pep=NO 61
married=NO children='(-inf-0.3]' 83
married=NO car=NO 102
married=NO car=YES 102
married=NO save_act=NO 67
married=NO save_act=YES 137
married=NO current_act=YES 162
married=NO mortgage=NO 130
married=NO mortgage=YES 74
married=NO pep=YES 120
married=NO pep=NO 84
married=YES children='(-inf-0.3]' 180
married=YES children='(0.9-1.2]' 89
married=YES children='(1.8-2.1]' 84
married=YES car=NO 202
married=YES car=YES 194
married=YES save_act=NO 119
married=YES save_act=YES 277
married=YES current_act=NO 103
married=YES current_act=YES 293
married=YES mortgage=NO 261
married=YES mortgage=YES 135
married=YES pep=YES 154
married=YES pep=NO 242
children='(-inf-0.3]' car=NO 139
children='(-inf-0.3]' car=YES 124

```
children='(-inf-0.3]' save_act=NO 89
children='(-inf-0.3]' save_act=YES 174
children='(-inf-0.3]' current_act=NO 64
children='(-inf-0.3]' current_act=YES 199
children='(-inf-0.3]' mortgage=NO 164
children='(-inf-0.3]' mortgage=YES 99
children='(-inf-0.3]' pep=YES 96
children='(-inf-0.3]' pep=NO 167
children='(0.9-1.2]' car=NO 68
children='(0.9-1.2]' car=YES 67
children='(0.9-1.2]' save_act=YES 95
children='(0.9-1.2]' current_act=YES 101
children='(0.9-1.2]' mortgage=NO 84
children='(0.9-1.2]' pep=YES 110
children='(1.8-2.1]' car=NO 63
children='(1.8-2.1]' car=YES 71
children='(1.8-2.1]' save_act=YES 99
children='(1.8-2.1]' current_act=YES 104
children='(1.8-2.1]' mortgage=NO 95
children='(1.8-2.1]' pep=NO 79
car=NO save_act=NO 99
car=NO save_act=YES 205
car=NO current_act=NO 69
car=NO current_act=YES 235
car=NO mortgage=NO 197
car=NO mortgage=YES 107
car=NO pep=YES 136
car=NO pep=NO 168
car=YES save_act=NO 87
car=YES save_act=YES 209
car=YES current_act=NO 76
car=YES current_act=YES 220
car=YES mortgage=NO 194
car=YES mortgage=YES 102
car=YES pep=YES 138
car=YES pep=NO 158
save_act=NO current_act=YES 136
save_act=NO mortgage=NO 121
save_act=NO mortgage=YES 65
save_act=NO pep=YES 95
save_act=NO pep=NO 91
save_act=YES current_act=NO 95
save_act=YES current_act=YES 319
save_act=YES mortgage=NO 270
save_act=YES mortgage=YES 144
save_act=YES pep=YES 179
save_act=YES pep=NO 235
current_act=NO mortgage=NO 90
current_act=NO pep=YES 63
current_act=NO pep=NO 82
current_act=YES mortgage=NO 301
current_act=YES mortgage=YES 154
current_act=YES pep=YES 211
current_act=YES pep=NO 244
mortgage=NO pep=YES 182
mortgage=NO pep=NO 209
```

mortgage=YES pep=YES 92
mortgage=YES pep=NO 117

Size of set of large itemsets L(3): 286

Large Itemsets L(3):
sex=FEMALE region=INNER_CITY married=YES 84
sex=FEMALE region=INNER_CITY children='(-inf-0.3]' 61
sex=FEMALE region=INNER_CITY car=NO 63
sex=FEMALE region=INNER_CITY car=YES 68
sex=FEMALE region=INNER_CITY save_act=YES 86
sex=FEMALE region=INNER_CITY current_act=YES 105
sex=FEMALE region=INNER_CITY mortgage=NO 88
sex=FEMALE region=INNER_CITY pep=NO 77
sex=FEMALE region=TOWN married=YES 67
sex=FEMALE region=TOWN save_act=YES 67
sex=FEMALE region=TOWN current_act=YES 63
sex=FEMALE region=TOWN mortgage=NO 60
sex=FEMALE married=NO save_act=YES 69
sex=FEMALE married=NO current_act=YES 84
sex=FEMALE married=NO mortgage=NO 67
sex=FEMALE married=NO pep=YES 62
sex=FEMALE married=YES children='(-inf-0.3]' 94
sex=FEMALE married=YES car=NO 99
sex=FEMALE married=YES car=YES 96
sex=FEMALE married=YES save_act=YES 137
sex=FEMALE married=YES current_act=YES 146
sex=FEMALE married=YES mortgage=NO 138
sex=FEMALE married=YES pep=YES 68
sex=FEMALE married=YES pep=NO 127
sex=FEMALE children='(-inf-0.3]' car=NO 68
sex=FEMALE children='(-inf-0.3]' car=YES 64
sex=FEMALE children='(-inf-0.3]' save_act=YES 88
sex=FEMALE children='(-inf-0.3]' current_act=YES 102
sex=FEMALE children='(-inf-0.3]' mortgage=NO 91
sex=FEMALE children='(-inf-0.3]' pep=NO 90
sex=FEMALE car=NO save_act=YES 106
sex=FEMALE car=NO current_act=YES 120
sex=FEMALE car=NO mortgage=NO 110
sex=FEMALE car=NO pep=YES 67
sex=FEMALE car=NO pep=NO 86
sex=FEMALE car=YES save_act=YES 100
sex=FEMALE car=YES current_act=YES 110
sex=FEMALE car=YES mortgage=NO 95
sex=FEMALE car=YES pep=YES 63
sex=FEMALE car=YES pep=NO 84
sex=FEMALE save_act=NO current_act=YES 70
sex=FEMALE save_act=NO mortgage=NO 67
sex=FEMALE save_act=YES current_act=YES 160
sex=FEMALE save_act=YES mortgage=NO 138
sex=FEMALE save_act=YES mortgage=YES 68
sex=FEMALE save_act=YES pep=YES 84
sex=FEMALE save_act=YES pep=NO 122
sex=FEMALE current_act=YES mortgage=NO 159
sex=FEMALE current_act=YES mortgage=YES 71
sex=FEMALE current_act=YES pep=YES 102

sex=FEMALE current_act=YES pep=NO 128
sex=FEMALE mortgage=NO pep=YES 90
sex=FEMALE mortgage=NO pep=NO 115
sex=MALE region=INNER_CITY married=YES 94
sex=MALE region=INNER_CITY children='(-inf-0.3]' 60
sex=MALE region=INNER_CITY car=NO 76
sex=MALE region=INNER_CITY car=YES 62
sex=MALE region=INNER_CITY save_act=YES 87
sex=MALE region=INNER_CITY current_act=YES 100
sex=MALE region=INNER_CITY mortgage=NO 87
sex=MALE region=INNER_CITY pep=YES 69
sex=MALE region=INNER_CITY pep=NO 69
sex=MALE region=TOWN save_act=YES 61
sex=MALE region=TOWN current_act=YES 65
sex=MALE married=NO save_act=YES 68
sex=MALE married=NO current_act=YES 78
sex=MALE married=NO mortgage=NO 63
sex=MALE married=YES children='(-inf-0.3]' 86
sex=MALE married=YES car=NO 103
sex=MALE married=YES car=YES 98
sex=MALE married=YES save_act=NO 61
sex=MALE married=YES save_act=YES 140
sex=MALE married=YES current_act=YES 147
sex=MALE married=YES mortgage=NO 123
sex=MALE married=YES mortgage=YES 78
sex=MALE married=YES pep=YES 86
sex=MALE married=YES pep=NO 115
sex=MALE children='(-inf-0.3]' car=NO 71
sex=MALE children='(-inf-0.3]' car=YES 60
sex=MALE children='(-inf-0.3]' save_act=YES 86
sex=MALE children='(-inf-0.3]' current_act=YES 97
sex=MALE children='(-inf-0.3]' mortgage=NO 73
sex=MALE children='(-inf-0.3]' pep=NO 77
sex=MALE car=NO save_act=YES 99
sex=MALE car=NO current_act=YES 115
sex=MALE car=NO mortgage=NO 87
sex=MALE car=NO mortgage=YES 64
sex=MALE car=NO pep=YES 69
sex=MALE car=NO pep=NO 82
sex=MALE car=YES save_act=YES 109
sex=MALE car=YES current_act=YES 110
sex=MALE car=YES mortgage=NO 99
sex=MALE car=YES pep=YES 75
sex=MALE car=YES pep=NO 74
sex=MALE save_act=NO current_act=YES 66
sex=MALE save_act=YES current_act=YES 159
sex=MALE save_act=YES mortgage=NO 132
sex=MALE save_act=YES mortgage=YES 76
sex=MALE save_act=YES pep=YES 95
sex=MALE save_act=YES pep=NO 113
sex=MALE current_act=YES mortgage=NO 142
sex=MALE current_act=YES mortgage=YES 83
sex=MALE current_act=YES pep=YES 109
sex=MALE current_act=YES pep=NO 116
sex=MALE mortgage=NO pep=YES 92
sex=MALE mortgage=NO pep=NO 94

sex=MALE mortgage=YES pep=NO 62
region=INNER_CITY married=NO current_act=YES 69
region=INNER_CITY married=YES children='(-inf-0.3]' 85
region=INNER_CITY married=YES car=NO 94
region=INNER_CITY married=YES car=YES 84
region=INNER_CITY married=YES save_act=YES 120
region=INNER_CITY married=YES current_act=YES 136
region=INNER_CITY married=YES mortgage=NO 116
region=INNER_CITY married=YES mortgage=YES 62
region=INNER_CITY married=YES pep=YES 66
region=INNER_CITY married=YES pep=NO 112
region=INNER_CITY children='(-inf-0.3]' car=NO 66
region=INNER_CITY children='(-inf-0.3]' save_act=YES 74
region=INNER_CITY children='(-inf-0.3]' current_act=YES 94
region=INNER_CITY children='(-inf-0.3]' mortgage=NO 79
region=INNER_CITY children='(-inf-0.3]' pep=NO 73
region=INNER_CITY car=NO save_act=YES 88
region=INNER_CITY car=NO current_act=YES 105
region=INNER_CITY car=NO mortgage=NO 91
region=INNER_CITY car=NO pep=YES 64
region=INNER_CITY car=NO pep=NO 75
region=INNER_CITY car=YES save_act=YES 85
region=INNER_CITY car=YES current_act=YES 100
region=INNER_CITY car=YES mortgage=NO 84
region=INNER_CITY car=YES pep=NO 71
region=INNER_CITY save_act=NO current_act=YES 69
region=INNER_CITY save_act=NO mortgage=NO 63
region=INNER_CITY save_act=YES current_act=YES 136
region=INNER_CITY save_act=YES mortgage=NO 112
region=INNER_CITY save_act=YES mortgage=YES 61
region=INNER_CITY save_act=YES pep=YES 73
region=INNER_CITY save_act=YES pep=NO 100
region=INNER_CITY current_act=YES mortgage=NO 136
region=INNER_CITY current_act=YES mortgage=YES 69
region=INNER_CITY current_act=YES pep=YES 90
region=INNER_CITY current_act=YES pep=NO 115
region=INNER_CITY mortgage=NO pep=YES 79
region=INNER_CITY mortgage=NO pep=NO 96
region=TOWN married=YES save_act=YES 86
region=TOWN married=YES current_act=YES 80
region=TOWN married=YES mortgage=NO 71
region=TOWN married=YES pep=NO 75
region=TOWN car=YES save_act=YES 70
region=TOWN car=YES current_act=YES 69
region=TOWN save_act=YES current_act=YES 94
region=TOWN save_act=YES mortgage=NO 79
region=TOWN save_act=YES pep=NO 76
region=TOWN current_act=YES mortgage=NO 79
region=TOWN current_act=YES pep=NO 74
married=NO children='(-inf-0.3]' current_act=YES 66
married=NO car=NO save_act=YES 72
married=NO car=NO current_act=YES 84
married=NO car=NO mortgage=NO 64
married=NO car=NO pep=YES 60
married=NO car=YES save_act=YES 65
married=NO car=YES current_act=YES 78

```
married=NO car=YES mortgage=NO 66
married=NO car=YES pep=YES 60
married=NO save_act=YES current_act=YES 113
married=NO save_act=YES mortgage=NO 86
married=NO save_act=YES pep=YES 77
married=NO save_act=YES pep=NO 60
married=NO current_act=YES mortgage=NO 102
married=NO current_act=YES mortgage=YES 60
married=NO current_act=YES pep=YES 95
married=NO current_act=YES pep=NO 67
married=NO mortgage=NO pep=YES 92
married=YES children='(-inf-0.3]' car=NO 100
married=YES children='(-inf-0.3]' car=YES 80
married=YES children='(-inf-0.3]' save_act=NO 61
married=YES children='(-inf-0.3]' save_act=YES 119
married=YES children='(-inf-0.3]' current_act=YES 133
married=YES children='(-inf-0.3]' mortgage=NO 116
married=YES children='(-inf-0.3]' mortgage=YES 64
married=YES children='(-inf-0.3]' pep=NO 141
married=YES children='(0.9-1.2]' save_act=YES 65
married=YES children='(0.9-1.2]' current_act=YES 65
married=YES children='(0.9-1.2]' pep=YES 74
married=YES children='(1.8-2.1]' save_act=YES 60
married=YES children='(1.8-2.1]' current_act=YES 62
married=YES car=NO save_act=NO 69
married=YES car=NO save_act=YES 133
married=YES car=NO current_act=YES 151
married=YES car=NO mortgage=NO 133
married=YES car=NO mortgage=YES 69
married=YES car=NO pep=YES 76
married=YES car=NO pep=NO 126
married=YES car=YES save_act=YES 144
married=YES car=YES current_act=YES 142
married=YES car=YES mortgage=NO 128
married=YES car=YES mortgage=YES 66
married=YES car=YES pep=YES 78
married=YES car=YES pep=NO 116
married=YES save_act=NO current_act=YES 87
married=YES save_act=NO mortgage=NO 77
married=YES save_act=NO pep=NO 67
married=YES save_act=YES current_act=NO 71
married=YES save_act=YES current_act=YES 206
married=YES save_act=YES mortgage=NO 184
married=YES save_act=YES mortgage=YES 93
married=YES save_act=YES pep=YES 102
married=YES save_act=YES pep=NO 175
married=YES current_act=NO mortgage=NO 62
married=YES current_act=NO pep=NO 65
married=YES current_act=YES mortgage=NO 199
married=YES current_act=YES mortgage=YES 94
married=YES current_act=YES pep=YES 116
married=YES current_act=YES pep=NO 177
married=YES mortgage=NO pep=YES 90
married=YES mortgage=NO pep=NO 171
married=YES mortgage=YES pep=YES 64
married=YES mortgage=YES pep=NO 71
```

children='(-inf-0.3]' car=NO save_act=YES 89
children='(-inf-0.3]' car=NO current_act=YES 107
children='(-inf-0.3]' car=NO mortgage=NO 92
children='(-inf-0.3]' car=NO pep=NO 91
children='(-inf-0.3]' car=YES save_act=YES 85
children='(-inf-0.3]' car=YES current_act=YES 92
children='(-inf-0.3]' car=YES mortgage=NO 72
children='(-inf-0.3]' car=YES pep=NO 76
children='(-inf-0.3]' save_act=NO current_act=YES 66
children='(-inf-0.3]' save_act=YES current_act=YES 133
children='(-inf-0.3]' save_act=YES mortgage=NO 112
children='(-inf-0.3]' save_act=YES mortgage=YES 62
children='(-inf-0.3]' save_act=YES pep=NO 131
children='(-inf-0.3]' current_act=YES mortgage=NO 125
children='(-inf-0.3]' current_act=YES mortgage=YES 74
children='(-inf-0.3]' current_act=YES pep=YES 72
children='(-inf-0.3]' current_act=YES pep=NO 127
children='(-inf-0.3]' mortgage=NO pep=NO 107
children='(-inf-0.3]' mortgage=YES pep=NO 60
children='(0.9-1.2]' save_act=YES current_act=YES 73
children='(0.9-1.2]' save_act=YES pep=YES 80
children='(0.9-1.2]' current_act=YES mortgage=NO 68
children='(0.9-1.2]' current_act=YES pep=YES 84
children='(0.9-1.2]' mortgage=NO pep=YES 71
children='(1.8-2.1]' save_act=YES current_act=YES 78
children='(1.8-2.1]' save_act=YES mortgage=NO 69
children='(1.8-2.1]' current_act=YES mortgage=NO 73
car=NO save_act=NO current_act=YES 76
car=NO save_act=NO mortgage=NO 68
car=NO save_act=YES current_act=YES 159
car=NO save_act=YES mortgage=NO 129
car=NO save_act=YES mortgage=YES 76
car=NO save_act=YES pep=YES 88
car=NO save_act=YES pep=NO 117
car=NO current_act=YES mortgage=NO 158
car=NO current_act=YES mortgage=YES 77
car=NO current_act=YES pep=YES 110
car=NO current_act=YES pep=NO 125
car=NO mortgage=NO pep=YES 89
car=NO mortgage=NO pep=NO 108
car=NO mortgage=YES pep=NO 60
car=YES save_act=NO current_act=YES 60
car=YES save_act=YES current_act=YES 160
car=YES save_act=YES mortgage=NO 141
car=YES save_act=YES mortgage=YES 68
car=YES save_act=YES pep=YES 91
car=YES save_act=YES pep=NO 118
car=YES current_act=YES mortgage=NO 143
car=YES current_act=YES mortgage=YES 77
car=YES current_act=YES pep=YES 101
car=YES current_act=YES pep=NO 119
car=YES mortgage=NO pep=YES 93
car=YES mortgage=NO pep=NO 101
save_act=NO current_act=YES mortgage=NO 89
save_act=NO current_act=YES pep=YES 71
save_act=NO current_act=YES pep=NO 65

```
sex=FEMALE region=INNER_CITY current_act=YES mortgage=NO 71
sex=FEMALE region=INNER_CITY current_act=YES pep=NO 64
sex=FEMALE married=YES children='(-inf-0.3]' save_act=YES 63
sex=FEMALE married=YES children='(-inf-0.3]' current_act=YES 71
sex=FEMALE married=YES children='(-inf-0.3]' mortgage=NO 70
sex=FEMALE married=YES children='(-inf-0.3]' pep=NO 78
sex=FEMALE married=YES car=NO save_act=YES 68
sex=FEMALE married=YES car=NO current_act=YES 76
sex=FEMALE married=YES car=NO mortgage=NO 76
sex=FEMALE married=YES car=NO pep=NO 66
sex=FEMALE married=YES car=YES save_act=YES 69
sex=FEMALE married=YES car=YES current_act=YES 70
sex=FEMALE married=YES car=YES mortgage=NO 62
sex=FEMALE married=YES car=YES pep=NO 61
sex=FEMALE married=YES save_act=YES current_act=YES 103
sex=FEMALE married=YES save_act=YES mortgage=NO 95
sex=FEMALE married=YES save_act=YES pep=NO 91
sex=FEMALE married=YES current_act=YES mortgage=NO 106
sex=FEMALE married=YES current_act=YES pep=NO 93
sex=FEMALE married=YES mortgage=NO pep=NO 93
sex=FEMALE children='(-inf-0.3]' save_act=YES current_act=YES 68
sex=FEMALE children='(-inf-0.3]' save_act=YES mortgage=NO 62
sex=FEMALE children='(-inf-0.3]' save_act=YES pep=NO 69
sex=FEMALE children='(-inf-0.3]' current_act=YES mortgage=NO 73
sex=FEMALE children='(-inf-0.3]' current_act=YES pep=NO 70
sex=FEMALE children='(-inf-0.3]' mortgage=NO pep=NO 64
sex=FEMALE car=NO save_act=YES current_act=YES 82
sex=FEMALE car=NO save_act=YES mortgage=NO 75
sex=FEMALE car=NO save_act=YES pep=NO 64
sex=FEMALE car=NO current_act=YES mortgage=NO 87
sex=FEMALE car=NO current_act=YES pep=NO 64
sex=FEMALE car=NO mortgage=NO pep=NO 64
sex=FEMALE car=YES save_act=YES current_act=YES 78
sex=FEMALE car=YES save_act=YES mortgage=NO 63
sex=FEMALE car=YES current_act=YES mortgage=NO 72
sex=FEMALE car=YES current_act=YES pep=NO 64
```

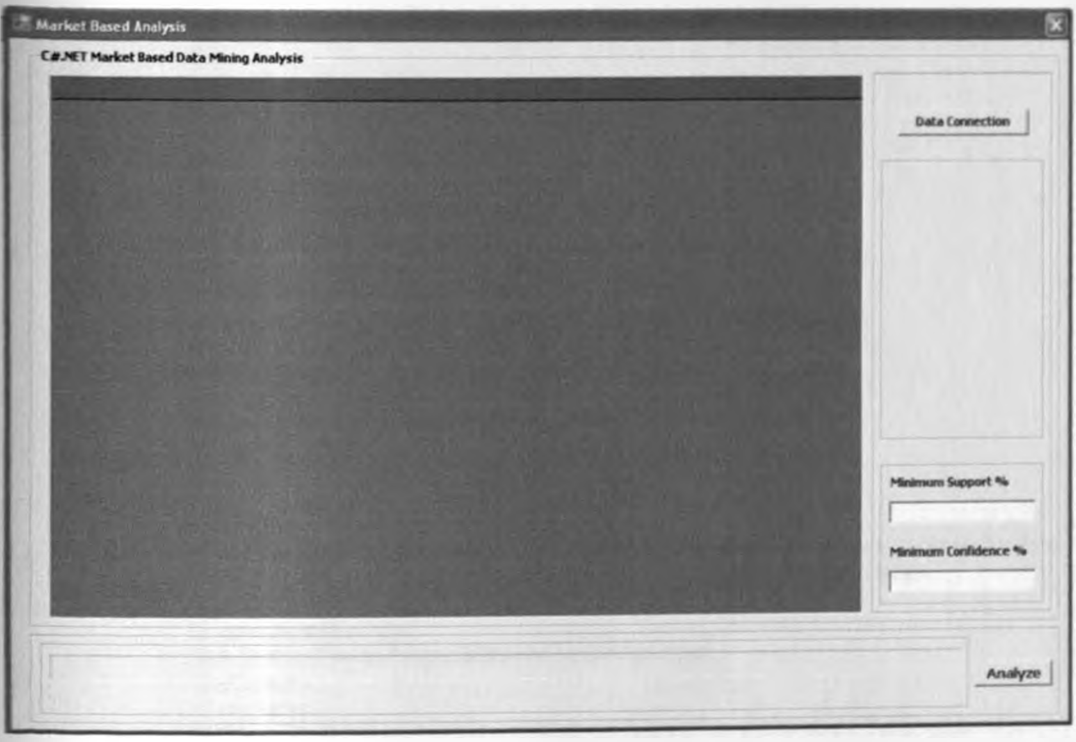
sex=FEMALE save_act=YES current_act=YES mortgage=NO 107
sex=FEMALE save_act=YES current_act=YES pep=YES 68
sex=FEMALE save_act=YES current_act=YES pep=NO 92
sex=FEMALE save_act=YES mortgage=NO pep=YES 61
sex=FEMALE save_act=YES mortgage=NO pep=NO 77
sex=FEMALE current_act=YES mortgage=NO pep=YES 73
sex=FEMALE current_act=YES mortgage=NO pep=NO 86
sex=MALE region=INNER_CITY married=YES save_act=YES 60
sex=MALE region=INNER_CITY married=YES current_act=YES 67
sex=MALE region=INNER_CITY save_act=YES current_act=YES 65
sex=MALE region=INNER_CITY current_act=YES mortgage=NO 65
sex=MALE married=YES children='(-inf-0.3]' current_act=YES 62
sex=MALE married=YES children='(-inf-0.3]' pep=NO 63
sex=MALE married=YES car=NO save_act=YES 65
sex=MALE married=YES car=NO current_act=YES 75
sex=MALE married=YES car=NO pep=NO 60
sex=MALE married=YES car=YES save_act=YES 75
sex=MALE married=YES car=YES current_act=YES 72
sex=MALE married=YES car=YES mortgage=NO 66
sex=MALE married=YES save_act=YES current_act=YES 103
sex=MALE married=YES save_act=YES mortgage=NO 89
sex=MALE married=YES save_act=YES pep=NO 84
sex=MALE married=YES current_act=YES mortgage=NO 93
sex=MALE married=YES current_act=YES pep=YES 63
sex=MALE married=YES current_act=YES pep=NO 84
sex=MALE married=YES mortgage=NO pep=NO 78
sex=MALE children='(-inf-0.3]' save_act=YES current_act=YES 65
sex=MALE children='(-inf-0.3]' save_act=YES pep=NO 62
sex=MALE car=NO save_act=YES current_act=YES 77
sex=MALE car=NO current_act=YES mortgage=NO 71
sex=MALE car=NO current_act=YES pep=NO 61
sex=MALE car=YES save_act=YES current_act=YES 82
sex=MALE car=YES save_act=YES mortgage=NO 78
sex=MALE car=YES save_act=YES pep=NO 60
sex=MALE car=YES current_act=YES mortgage=NO 71
sex=MALE save_act=YES current_act=YES mortgage=NO 105
sex=MALE save_act=YES current_act=YES pep=YES 72
sex=MALE save_act=YES current_act=YES pep=NO 87
sex=MALE save_act=YES mortgage=NO pep=YES 67
sex=MALE save_act=YES mortgage=NO pep=NO 65
sex=MALE current_act=YES mortgage=NO pep=YES 70
sex=MALE current_act=YES mortgage=NO pep=NO 72
region=INNER_CITY married=YES children='(-inf-0.3]' current_act=YES 68
region=INNER_CITY married=YES children='(-inf-0.3]' pep=NO 64
region=INNER_CITY married=YES car=NO save_act=YES 60
region=INNER_CITY married=YES car=NO current_act=YES 69
region=INNER_CITY married=YES car=NO mortgage=NO 62
region=INNER_CITY married=YES car=YES save_act=YES 60
region=INNER_CITY married=YES car=YES current_act=YES 67
region=INNER_CITY married=YES save_act=YES current_act=YES 94
region=INNER_CITY married=YES save_act=YES mortgage=NO 78
region=INNER_CITY married=YES save_act=YES pep=NO 80
region=INNER_CITY married=YES current_act=YES mortgage=NO 92
region=INNER_CITY married=YES current_act=YES pep=NO 89
region=INNER_CITY married=YES mortgage=NO pep=NO 80
region=INNER_CITY children='(-inf-0.3]' save_act=YES current_act=YES 61

```
region=INNER_CITY children='(-inf-0.3]' current_act=YES mortgage=NO 61
region=INNER_CITY children='(-inf-0.3]' current_act=YES pep=NO 61
region=INNER_CITY car=NO save_act=YES current_act=YES 69
region=INNER_CITY car=NO current_act=YES mortgage=NO 73
region=INNER_CITY car=YES save_act=YES current_act=YES 67
region=INNER_CITY car=YES current_act=YES mortgage=NO 63
region=INNER_CITY save_act=YES current_act=YES mortgage=NO 90
region=INNER_CITY save_act=YES current_act=YES pep=NO 80
region=INNER_CITY save_act=YES mortgage=NO pep=NO 61
region=INNER_CITY current_act=YES mortgage=NO pep=NO 78
married=NO car=NO save_act=YES current_act=YES 60
married=NO save_act=YES current_act=YES mortgage=NO 70
married=NO save_act=YES current_act=YES pep=YES 64
married=NO save_act=YES mortgage=NO pep=YES 64
married=NO current_act=YES mortgage=NO pep=YES 73
married=YES children='(-inf-0.3]' car=NO save_act=YES 64
married=YES children='(-inf-0.3]' car=NO current_act=YES 74
married=YES children='(-inf-0.3]' car=NO mortgage=NO 67
married=YES children='(-inf-0.3]' car=NO pep=NO 80
married=YES children='(-inf-0.3]' car=YES pep=NO 61
married=YES children='(-inf-0.3]' save_act=YES current_act=YES 87
married=YES children='(-inf-0.3]' save_act=YES mortgage=NO 80
married=YES children='(-inf-0.3]' save_act=YES pep=NO 107
married=YES children='(-inf-0.3]' current_act=YES mortgage=NO 88
married=YES children='(-inf-0.3]' current_act=YES pep=NO 105
married=YES children='(-inf-0.3]' mortgage=NO pep=NO 104
married=YES car=NO save_act=YES current_act=YES 99
married=YES car=NO save_act=YES mortgage=NO 84
married=YES car=NO save_act=YES pep=NO 87
married=YES car=NO current_act=YES mortgage=NO 104
married=YES car=NO current_act=YES pep=NO 92
married=YES car=NO mortgage=NO pep=NO 89
married=YES car=YES save_act=YES current_act=YES 107
married=YES car=YES save_act=YES mortgage=NO 100
married=YES car=YES save_act=YES pep=NO 88
married=YES car=YES current_act=YES mortgage=NO 95
married=YES car=YES current_act=YES pep=NO 85
married=YES car=YES mortgage=NO pep=NO 82
married=YES save_act=YES current_act=YES mortgage=NO 142
married=YES save_act=YES current_act=YES mortgage=YES 64
married=YES save_act=YES current_act=YES pep=YES 76
married=YES save_act=YES current_act=YES pep=NO 130
married=YES save_act=YES mortgage=NO pep=YES 64
married=YES save_act=YES mortgage=NO pep=NO 120
married=YES current_act=YES mortgage=NO pep=YES 70
married=YES current_act=YES mortgage=NO pep=NO 129
children='(-inf-0.3]' car=NO save_act=YES current_act=YES 70
children='(-inf-0.3]' car=NO save_act=YES pep=NO 67
children='(-inf-0.3]' car=NO current_act=YES mortgage=NO 72
children='(-inf-0.3]' car=NO current_act=YES pep=NO 69
children='(-inf-0.3]' car=NO mortgage=NO pep=NO 62
children='(-inf-0.3]' car=YES save_act=YES current_act=YES 63
children='(-inf-0.3]' car=YES save_act=YES pep=NO 64
children='(-inf-0.3]' save_act=YES current_act=YES mortgage=NO 87
children='(-inf-0.3]' save_act=YES current_act=YES pep=NO 101
children='(-inf-0.3]' save_act=YES mortgage=NO pep=NO 74
```

```
children='(-inf-0.3]' current_act=YES mortgage=NO pep=NO 82
children='(0.9-1.2]' save_act=YES current_act=YES pep=YES 63
car=NO save_act=YES current_act=YES mortgage=NO 104
car=NO save_act=YES current_act=YES pep=YES 72
car=NO save_act=YES current_act=YES pep=NO 87
car=NO save_act=YES mortgage=NO pep=YES 61
car=NO save_act=YES mortgage=NO pep=NO 68
car=NO current_act=YES mortgage=NO pep=YES 77
car=NO current_act=YES mortgage=NO pep=NO 81
car=YES save_act=YES current_act=YES mortgage=NO 108
car=YES save_act=YES current_act=YES pep=YES 68
car=YES save_act=YES current_act=YES pep=NO 92
car=YES save_act=YES mortgage=NO pep=YES 67
car=YES save_act=YES mortgage=NO pep=NO 74
car=YES current_act=YES mortgage=NO pep=YES 66
car=YES current_act=YES mortgage=NO pep=NO 77
save_act=YES current_act=YES mortgage=NO pep=YES 104
save_act=YES current_act=YES mortgage=NO pep=NO 108
save_act=YES current_act=YES mortgage=YES pep=NO 71

Size of set of large itemsets L(5): 26

Large Itemsets L(5):
sex=FEMALE married=YES children='(-inf-0.3]' current_act=YES pep=NO 60
sex=FEMALE married=YES children='(-inf-0.3]' mortgage=NO pep=NO 63
sex=FEMALE married=YES car=NO current_act=YES mortgage=NO 60
sex=FEMALE married=YES save_act=YES current_act=YES mortgage=NO 72
sex=FEMALE married=YES save_act=YES current_act=YES pep=NO 67
sex=FEMALE married=YES save_act=YES mortgage=NO pep=NO 64
sex=FEMALE married=YES current_act=YES mortgage=NO pep=NO 70
sex=MALE married=YES save_act=YES current_act=YES mortgage=NO 70
sex=MALE married=YES save_act=YES current_act=YES pep=NO 63
region=INNER_CITY married=YES save_act=YES current_act=YES mortgage=NO 63
region=INNER_CITY married=YES save_act=YES current_act=YES pep=NO 65
region=INNER_CITY married=YES current_act=YES mortgage=NO pep=NO 66
married=YES children='(-inf-0.3]' car=NO current_act=YES pep=NO 60
married=YES children='(-inf-0.3]' car=NO mortgage=NO pep=NO 60
married=YES children='(-inf-0.3]' save_act=YES current_act=YES mortgage=NO 61
married=YES children='(-inf-0.3]' save_act=YES current_act=YES pep=NO 80
married=YES children='(-inf-0.3]' save_act=YES mortgage=NO pep=NO 73
married=YES children='(-inf-0.3]' current_act=YES mortgage=NO pep=NO 80
married=YES car=NO save_act=YES current_act=YES mortgage=NO 66
married=YES car=NO save_act=YES current_act=YES pep=NO 64
married=YES car=NO current_act=YES mortgage=NO pep=NO 66
married=YES car=YES save_act=YES current_act=YES mortgage=NO 76
married=YES car=YES save_act=YES current_act=YES pep=NO 66
married=YES car=YES save_act=YES mortgage=NO pep=NO 63
married=YES car=YES current_act=YES mortgage=NO pep=NO 63
married=YES save_act=YES current_act=YES mortgage=NO pep=NO 91

Best rules found:

 1. children='(-inf-0.3]' save_act=YES mortgage=NO pep=NO 74 ==> married=YES 73
conf:(0.99)
 2. sex=FEMALE children='(-inf-0.3]' mortgage=NO pep=NO 64 ==> married=YES 63
conf:(0.98)
```

3. children='(-inf-0.3]' current_act=YES mortgage=NO pep=NO 82 ==> married=YES 80 conf:(0.98)
4. children='(-inf-0.3]' mortgage=NO pep=NO 107 ==> married=YES 104 conf:(0.97)
5. children='(-inf-0.3]' car=NO mortgage=NO pep=NO 62 ==> married=YES 60 conf:(0.97)
6. married=YES children='(-inf-0.3]' save_act=YES current_act=YES 87 ==> pep=NO 80 conf:(0.92)
7. married=YES children='(-inf-0.3]' save_act=YES mortgage=NO 80 ==> pep=NO 73 conf:(0.91)
8. married=YES children='(-inf-0.3]' current_act=YES mortgage=NO 88 ==> pep=NO 80 conf:(0.91)
9. sex=FEMALE married=YES children='(-inf-0.3]' mortgage=NO 70 ==> pep=NO 63 conf:(0.9)

4.0 Source Code – *Apriori* Algorithm




```
using System;
using System.Drawing;
using System.Collections;
using System.ComponentModel;
using System.Windows.Forms;
using System.Data;
using VISUAL_BASIC_DATA_MINING_NET;
using VISUAL_BASIC_DATA_MINING_NET.CustomEvents;
using VISUAL_BASIC_DATA_MINING_NET.DataTransformationServices;


namespace APrioriWindows
{
	/// <summary>
	/// Summary description for MarketBasedAnalysis.
	/// </summary>
	public class MarketBasedAnalysis : System.Windows.Forms.Form
	{
		private System.Windows.Forms.GroupBox groupBox1;
		private System.Windows.Forms.DataGrid dataGridViewAnalysisResult;
		private System.Windows.Forms.GroupBox groupBoxCommands;
		private System.Windows.Forms.Button buttonDataConnection;
		//
		//
		private ConnectionDialogBox connectionDialogBox;
		private System.Windows.Forms.Button buttonOK;
		//
		//
		private System.Windows.Forms.GroupBox groupBoxSettings;
```

```
private System.Windows.Forms.Label lblSupportCount;
private System.Windows.Forms.Label lblMinimumConfidence;
private System.Windows.Forms.TextBox txtMinimumSupport;
private System.Windows.Forms.TextBox txtMinimumConfidence;
//
//
private DataMining DMS;
private ViewData dataView;
private Data dataAnalysis;
private NorthwindDTS dts;
private Data orders;
private string minimumConfidence;
private string minimumSupport;
private int minimumConfidenceLength;
private int minimumSupportLength;
private System.Windows.Forms.GroupBox groupBoxProgressMonitor;
private System.Windows.Forms.GroupBox groupBox2;
private System.Windows.Forms.Label lblProgressBar;
private System.Windows.Forms.GroupBox groupBoxViewTables;
private System.Windows.Forms.ProgressBar progressBarMonitor;
//
//
//
/// <summary>
/// The public OnProgressMonitorEvent raises the ProgressMonitorEvent event by invoking
/// the delegates. The sender is always this, the current instance of the class.
/// </summary>
/// <param name="e">
/// A CustomEvents.ProgressMonitorEventArgs object.
/// </param>
/// <remarks>
/// This method is used to invoke a dalegate that notifies clients about the progress of an
executing code.
/// </remarks>
public void OnProgressMonitorEvent(object sender, ProgressMonitorEventArgs e)
{
        //Sets the information to be displayed on the progress bar
        this.progressBarMonitor.Minimum = e.MinimumValue;

        this.progressBarMonitor.Maximum = e.MaximumValue;

        this.progressBarMonitor.Value = e.CurrentValue;

        this.progressBarMonitor.Refresh();

        this.lblProgressBar.Text = e.EventMessage;

        this.lblProgressBar.Refresh();
}


/// <summary>
/// A custom event that notifies clients about the progress of the executing code.
/// </summary>
public event ProgressMonitorEventHandler ProgressMonitorEvent;


//
/// <summary>
/// Required designer variable.
/// </summary>
private System.ComponentModel.Container components = null;
```

```
public MarketBasedAnalysis()
{
	//
	// Required for Windows Form Designer support
	//
	InitializeComponent();

	//
	// TODO: Add any constructor code after InitializeComponent call
	//
}

/// <summary>
/// Clean up any resources being used.
/// </summary>
protected override void Dispose( bool disposing )
{
	if( disposing )
	{
		if(components != null)
		{
			components.Dispose();
		}
	}
	base.Dispose( disposing );
}

#region Windows Form Designer generated code
/// <summary>
/// Required method for Designer support - do not modify
/// the contents of this method with the code editor.
/// </summary>
private void InitializeComponent()
{
	this.groupBox1 = new System.Windows.Forms.GroupBox();
	this.groupBoxCommands = new System.Windows.Forms.GroupBox();
	this.groupBoxSettings = new System.Windows.Forms.GroupBox();
	this.lblMinimumConfidence = new System.Windows.Forms.Label();
	this.txtMinimumConfidence = new System.Windows.Forms.TextBox();
	this.lblSupportCount = new System.Windows.Forms.Label();
	this.txtMinimumSupport = new System.Windows.Forms.TextBox();
	this.buttonDataConnection = new System.Windows.Forms.Button();
	this.dataGridViewAnalysisResult = new System.Windows.Forms.DataGrid();
	this.groupBoxProgressMonitor = new System.Windows.Forms.GroupBox();
	this.groupBox2 = new System.Windows.Forms.GroupBox();
	this.lblProgressBar = new System.Windows.Forms.Label();
	this.progressBarMonitor = new System.Windows.Forms.ProgressBar();
	this.buttonOK = new System.Windows.Forms.Button();
	this.groupBoxViewTables = new System.Windows.Forms.GroupBox();
	this.groupBox1.SuspendLayout();
	this.groupBoxCommands.SuspendLayout();
	this.groupBoxSettings.SuspendLayout();

((System.ComponentModel.ISupportInitialize)(this.dataGridViewAnalysisResult)).BeginInit();
	this.groupBoxProgressMonitor.SuspendLayout();
	this.groupBox2.SuspendLayout();
	this.SuspendLayout();
	//
	// groupBox1
	//
	this.groupBox1.Controls.AddRange(new System.Windows.Forms.Control[] {
									this.groupBoxCommands,
```

```
this.dataGridViewAnalysisResult});
			this.groupBox1.Font = new System.Drawing.Font("Tahoma", 8.25F,
System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
			this.groupBox1.Location = new System.Drawing.Point(16, 8);
			this.groupBox1.Name = "groupBox1";
			this.groupBox1.Size = new System.Drawing.Size(976, 528);
			this.groupBox1.TabIndex = 0;
			this.groupBox1.TabStop = false;
			this.groupBox1.Text = "C#.NET Market Based Data Mining Analysis";
			//
			// groupBoxCommands
			//
			this.groupBoxCommands.Controls.AddRange(new
System.Windows.Forms.Control[] {

this.groupBoxViewTables,

this.groupBoxSettings,

this.buttonDataConnection});
			this.groupBoxCommands.Location = new System.Drawing.Point(792, 16);
			this.groupBoxCommands.Name = "groupBoxCommands";
			this.groupBoxCommands.Size = new System.Drawing.Size(176, 504);
			this.groupBoxCommands.TabIndex = 2;
			this.groupBoxCommands.TabStop = false;
			//
			// groupBoxSettings
			//
			this.groupBoxSettings.Controls.AddRange(new System.Windows.Forms.Control[]
{

this.lblMinimumConfidence,

this.txtMinimumConfidence,

this.lblSupportCount,

this.txtMinimumSupport});
			this.groupBoxSettings.Location = new System.Drawing.Point(8, 360);
			this.groupBoxSettings.Name = "groupBoxSettings";
			this.groupBoxSettings.Size = new System.Drawing.Size(160, 136);
			this.groupBoxSettings.TabIndex = 1;
			this.groupBoxSettings.TabStop = false;
			//
			// lblMinimumConfidence
			//
			this.lblMinimumConfidence.Font = new System.Drawing.Font("Tahoma", 8.25F,
System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
			this.lblMinimumConfidence.ForeColor = System.Drawing.Color.DarkBlue;
			this.lblMinimumConfidence.Location = new System.Drawing.Point(8, 80);
			this.lblMinimumConfidence.Name = "lblMinimumConfidence";
			this.lblMinimumConfidence.Size = new System.Drawing.Size(144, 16);
			this.lblMinimumConfidence.TabIndex = 3;
			this.lblMinimumConfidence.Text = "Minimum Confidence %";
```

```
            //
            // txtMinimumConfidence
            //
            this.txtMinimumConfidence.Location = new System.Drawing.Point(8, 104);
            this.txtMinimumConfidence.Name = "txtMinimumConfidence";
            this.txtMinimumConfidence.Size = new System.Drawing.Size(144, 21);
            this.txtMinimumConfidence.TabIndex = 2;
            this.txtMinimumConfidence.Text = "";
            this.txtMinimumConfidence.Validating += new
System.ComponentModel.CancelEventHandler(this.txtMinimumConfidence_Validating);
            //
            // lblSupportCount
            //
            this.lblSupportCount.Font = new System.Drawing.Font("Tahoma", 8.25F,
System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
            this.lblSupportCount.ForeColor = System.Drawing.Color.DarkBlue;
            this.lblSupportCount.Location = new System.Drawing.Point(8, 16);
            this.lblSupportCount.Name = "lblSupportCount";
            this.lblSupportCount.Size = new System.Drawing.Size(136, 16);
            this.lblSupportCount.TabIndex = 1;
            this.lblSupportCount.Text = "Minimum Support %";
            //
            // txtMinimumSupport
            //
            this.txtMinimumSupport.Location = new System.Drawing.Point(8, 40);
            this.txtMinimumSupport.Name = "txtMinimumSupport";
            this.txtMinimumSupport.Size = new System.Drawing.Size(144, 21);
            this.txtMinimumSupport.TabIndex = 0;
            this.txtMinimumSupport.Text = "";
            this.txtMinimumSupport.Validating += new
System.ComponentModel.CancelEventHandler(this.txtMinimumSupport_Validating);
            this.txtMinimumSupport.TextChanged += new
System.EventHandler(this.txtMinimumSupport_TextChanged);
            //
            // buttonDataConnection
            //
            this.buttonDataConnection.Font = new System.Drawing.Font("Tahoma", 8.25F,
System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
            this.buttonDataConnection.ForeColor = System.Drawing.Color.DarkBlue;
            this.buttonDataConnection.Location = new System.Drawing.Point(24, 40);
            this.buttonDataConnection.Name = "buttonDataConnection";
            this.buttonDataConnection.Size = new System.Drawing.Size(128, 24);
            this.buttonDataConnection.TabIndex = 0;
            this.buttonDataConnection.Text = "&Data Connection";
            this.buttonDataConnection.Click += new
System.EventHandler(this.buttonDataConnection_Click);
            //
            // dataGridViewAnalysisResult
            //
            this.dataGridViewAnalysisResult.AlternatingBackColor =
System.Drawing.Color.Gainsboro;
            this.dataGridViewAnalysisResult.DataMember = "";
            this.dataGridViewAnalysisResult.Font = new System.Drawing.Font("Tahoma",
8.25F, System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
            this.dataGridViewAnalysisResult.HeaderForeColor =
System.Drawing.SystemColors.ControlText;
            this.dataGridViewAnalysisResult.Location = new System.Drawing.Point(16, 24);
            this.dataGridViewAnalysisResult.Name = "dataGridViewAnalysisResult";
            this.dataGridViewAnalysisResult.ReadOnly = true;
            this.dataGridViewAnalysisResult.Size = new System.Drawing.Size(768, 496);
            this.dataGridViewAnalysisResult.TabIndex = 1;
            this.dataGridViewAnalysisResult.Navigate += new
System.Windows.Forms.NavigateEventHandler(this.dataGridViewAnalysisResult_Navigate);
```

```
//
// groupBoxProgressMonitor
//
this.groupBoxProgressMonitor.Controls.AddRange(new
System.Windows.Forms.Control[] {

                                        this.groupBox2,

this.buttonOK});
this.groupBoxProgressMonitor.Location = new System.Drawing.Point(16, 536);
this.groupBoxProgressMonitor.Name = "groupBoxProgressMonitor";
this.groupBoxProgressMonitor.Size = new System.Drawing.Size(976, 88);
this.groupBoxProgressMonitor.TabIndex = 1;
this.groupBoxProgressMonitor.TabStop = false;
//
// groupBox2
//
this.groupBox2.Controls.AddRange(new System.Windows.Forms.Control[] {

                                        this.lblProgressBar,

                                        this.progressBarMonitor});
this.groupBox2.Location = new System.Drawing.Point(8, 8);
this.groupBox2.Name = "groupBox2";
this.groupBox2.Size = new System.Drawing.Size(880, 72);
this.groupBox2.TabIndex = 2;
this.groupBox2.TabStop = false;
//
// lblProgressBar
//
this.lblProgressBar.Location = new System.Drawing.Point(16, 48);
this.lblProgressBar.Name = "lblProgressBar";
this.lblProgressBar.Size = new System.Drawing.Size(856, 16);
this.lblProgressBar.TabIndex = 1;
//
// progressBarMonitor
//
this.progressBarMonitor.Location = new System.Drawing.Point(8, 16);
this.progressBarMonitor.Name = "progressBarMonitor";
this.progressBarMonitor.Size = new System.Drawing.Size(864, 24);
this.progressBarMonitor.TabIndex = 0;
//
// buttonOK
//
this.buttonOK.Font = new System.Drawing.Font("Tahoma", 9.75F,
System.Drawing.FontStyle.Bold, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
this.buttonOK.ForeColor = System.Drawing.Color.DarkBlue;
this.buttonOK.Location = new System.Drawing.Point(893, 37);
this.buttonOK.Name = "buttonOK";
this.buttonOK.TabIndex = 1;
this.buttonOK.Text = "&Analyze";
this.buttonOK.Click += new System.EventHandler(this.buttonOK_Click);
//
// groupBoxViewTables
//
this.groupBoxViewTables.Location = new System.Drawing.Point(8, 80);
this.groupBoxViewTables.Name = "groupBoxViewTables";
this.groupBoxViewTables.Size = new System.Drawing.Size(160, 264);
this.groupBoxViewTables.TabIndex = 2;
this.groupBoxViewTables.TabStop = false;
//
// MarketBasedAnalysis
```

```
			//
			this.AutoScaleBaseSize = new System.Drawing.Size(5, 14);
			this.ClientSize = new System.Drawing.Size(1000, 629);
			this.Controls.AddRange(new System.Windows.Forms.Control[] {

								this.groupBoxProgressMonitor,

								this.groupBox1});
			this.Font = new System.Drawing.Font("Tahoma", 8.25F,
System.Drawing.FontStyle.Regular, System.Drawing.GraphicsUnit.Point, ((System.Byte)(0)));
			this.MaximizeBox = false;
			this.MinimizeBox = false;
			this.Name = "MarketBasedAnalysis";
			this.Text = "Market Based Analysis";
			this.TopMost = true;
			this.Load += new System.EventHandler(this.MarketBasedAnalysis_Load);
			this.groupBox1.ResumeLayout(false);
			this.groupBoxCommands.ResumeLayout(false);
			this.groupBoxSettings.ResumeLayout(false);

	((System.ComponentModel.ISupportInitialize)(this.dataGridViewAnalysisResult)).EndInit();
			this.groupBoxProgressMonitor.ResumeLayout(false);
			this.groupBox2.ResumeLayout(false);
			this.ResumeLayout(false);

		}
		#endregion

		/// <summary>
		/// The main entry point for this application.
		/// </summary>
		[STAThread]
		static void Main()
		{
			Application.Run(new MarketBasedAnalysis());
		}

		private void MarketBasedAnalysis_Load(object sender, System.EventArgs e)
		{

		}

		private void buttonDataConnection_Click(object sender, System.EventArgs e)
		{
			connectionDialogBox = new ConnectionDialogBox();

			connectionDialogBox.ShowDialog(this);
		}


		private void buttonOK_Click(object sender, System.EventArgs e)
		{
			this.DMS = new DataMining();


			this.DMS.ProgressMonitorEvent += new
ProgressMonitorEventHandler(this.OnProgressMonitorEvent);


			if((ClassInfo.DataStorage == ClassInfo.DataStorageLocation.Database) &&
(ClassInfo.DataStorageModel == ClassInfo.DataModel.TransactionsTable))
			{
				this.dataAnalysis =
```

```
DMS.MarketBasedAnalysis(ClassInfo.MinimumSupport, ClassInfo.MinimumConfidence,ClassInfo.

        ConnectionString,"TransactionsTable",CommandType.TableDirect);
                            }

                            else if (ClassInfo.DataStorage == ClassInfo.DataStorageLocation.XMLFile)
                            {
                                    this.dataAnalysis =
DMS.MarketBasedAnalysis(ClassInfo.MinimumSupport, ClassInfo.MinimumConfidence,ClassInfo.

                            XMLFilePath);
                            }


                            else if((ClassInfo.DataStorage == ClassInfo.DataStorageLocation.Database) &&
(ClassInfo.DataStorageModel == ClassInfo.DataModel.NorthwindDatabase ))
                            {
                                    dts = new NorthwindDTS();

                                    dts.LoadNorthwindwindProducts(ClassInfo.ConnectionString);

                                    dts.LoadNorthwindwindOrders(ClassInfo.ConnectionString);

                                    dts.LoadNorthwindwindOrderDetails(ClassInfo.ConnectionString);

                                    orders = dts.GetOrders();

                                    this.dataAnalysis =
DMS.MarketBasedAnalysis(ClassInfo.MinimumSupport,ClassInfo.MinimumConfidence,orders);
                            }


                            if(this.dataAnalysis != null)
                            {

                                    this.dataView = new ViewData();


        this.dataAnalysis.Tables.Add(this.dataView.CreateViewRulesTable(ClassInfo.MinimumConfidence,
this.dataAnalysis).Copy());


        this.dataAnalysis.Tables.Add(this.dataView.CreateViewSubsetTable(this.dataAnalysis).Copy());


                                    this.dataGridViewAnalysisResult.DataSource =
DMS.ViewDataMiningAnalysis("ViewRulesTable","Confidence desc");
                            }

                }



        private void dataGridViewAnalysisResult_Navigate(object sender,
System.Windows.Forms.NavigateEventArgs ne)
                {
                }

                private void txtMinimumSupport_TextChanged(object sender, System.EventArgs e)
```

```
                {
                }


                private void txtMinimumSupport_Validating(object sender,
System.ComponentModel.CancelEventArgs e)
                {
                        if (txtMinimumSupport.Text.Length == 0)
                        {
                                MessageBox.Show(this,"Please enter a minimum support between 0%
and 100%",ClassInfo.AppCaption,MessageBoxButtons.

                                        OK ,MessageBoxIcon.Information);
                        }

                        else if (txtMinimumSupport.Text.Length > 0)
                        {
                                minimumSupport = txtMinimumSupport.Text.Trim();

                                minimumSupportLength = minimumSupport.Length;

                                if (minimumSupport.EndsWith("%"))
                                {
                                        ClassInfo.MinimumSupport =
Convert.ToDouble(minimumSupport.Substring(0,

                                minimumSupportLength-1));
                                }
                                else
                                {
                                        ClassInfo.MinimumSupport =
Convert.ToDouble(minimumSupport.Substring(0,

                                minimumSupportLength));
                                }
                        }
                }


                private void txtMinimumConfidence_Validating(object sender,
System.ComponentModel.CancelEventArgs e)
                {
                        if (txtMinimumConfidence.Text.Length == 0)
                        {
                                MessageBox.Show(this,"Please enter a minimum confidence between
0% and 100%",ClassInfo.AppCaption,

                MessageBoxButtons.OK ,MessageBoxIcon.Information);
                        }

                        else if (txtMinimumConfidence.Text.Length > 0)
                        {
                                minimumConfidence = txtMinimumConfidence.Text.Trim();

                                minimumConfidenceLength = minimumConfidence.Length;

                                if (minimumConfidence.EndsWith("%"))
                                {

                                        ClassInfo.MinimumConfidence =
Convert.ToDouble(minimumConfidence.Substring(0,

                                minimumConfidenceLength-1));
```

```
                        }
                        else
                        {
                                ClassInfo.MinimumConfidence =
Convert.ToDouble(minimumConfidence.Substring(0,

                        minimumConfidenceLength));
                        }
                    }
                }



        }
}
```