Derivation and test of spatial rules for the prediction of efficiency of business branches

Ara Toomanian

March, 2007

Derivation and test of spatial rules for the prediction of efficiency of business branches

by

Ara Toomanian

Thesis submitted to the International Institute for Geo-information Science and Earth Observation in partial fulfilment of the requirements for the degree of Master of Science in Geo-information Science and Earth Observation, Specialisation: *Geoinformatics*.

Thesis Assessment Board

Thesis advisor ITC	Dr.Ir. R. A. de By W. H. Bakker, M.Sc.
Thesis advisor KNTU	Dr. A. Mansourian
Assessment Board chair	Prof.Dr. M. J. Kraak
Thesis examiners	Dr.Ir. M. van Keulen Dr. S. Mesgari

INTERNATIONAL INSTITUTE FOR GEO-INFORMATION SCIENCE AND EARTH OBSERVATION

ENSCHEDE, THE NETHERLANDS

K.N.T.U

K. N. TOOSI UNIVERSITY OF TECHNOLOGY

TEHRAN, IRAN

Disclaimer

This document describes work undertaken as part of a programme of study at the International Institute for Geo-information Science and Earth Observation (ITC). All views and opinions expressed therein remain the sole responsibility of the author, and do not necessarily represent those of the institute.

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, we use data mining algorithms to find and extract useful knowledge not only to help managers making better decisions for site selection, but also for extracting associations between parameters at different scales. We also tried to find a link between a mathematically determined efficiency measure and spatial association rules, which is a database method in data mining.

During the research, we classified the study area into three different classes as 'high', 'average', and 'low' according to the efficiency and turnover measures. Afterwards, in each class we used an a priori like algorithm to find the most frequent item sets and predict an average range of efficiency. For the second scenario we also put a limitation in the a priori like algorithm to derive a constraint-based frequent item set containing the efficiency measure parameters.

In general, as the efficiency measure in the low class had a higher frequency than in other classes, we obtained negative rules 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 we recommend to use normal distributed efficiency measure data 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. Needless to say, such models can be used for the parameters inside the frequent item set.

Keywords

Spatial Association Rule, DEA, Efficiency, A priori, Multiple scale

Acknowledgements

I would like to take the opportunity to express my deepest gratitude to all the people who support and help me in the study and research work.

I would like to express my gratitude to my father, Prof.Dr. Megerdich Toomanian, and my mother, Mrs. Armenoohi Sarkisians, for supporting me in the whole period of my life. Without their unlimited help, I could not finish my research.

I am deeply indebted to my lovely wife, Mrs. Hripsimeh Azmikaelians, and her family for their cooperation and support. I will never forget her positive influence on my work during the period of M.Sc. research.

I also would like to appreciate my sister Mrs.Dr. Marina Toomanian, and her husband Mr. Saro Avakians, and my cute nephew Mr. Aren Avakians, for their kindness during the time of my study.

I would like to express my deep and sincere gratitude to my first supervisor, Dr.Ir. R. A. de By, for opening a new window in my scientific life. I appreciate him for improving and shaping my research skills. Without his encouragement, I really could not finish my research. During the research period, we had very useful scientific weekly discussions. Honestly, I was counting the days to meet him every week!. I greatly admire his style of supervision.

I am thankful to my second supervisor W. H. Bakker, M.Sc. for his valuable comments on my thesis. His useful guidelines and ideas were really helpful.

I profess my thanks and regards to my Iranian supervisor Dr. A. Mansourian for his comments on the thesis in the early stage of the research.

I would like to express my gratitude to ITC and KNTU for giving me the opportunity to complete my MSc thesis. I also extend my profound gratitude to M.Sc. G. Huurneman and Dr. B. Vosooghi the program directors of GFM/JKIP, for establishment of the JKIP joint program.

I am also grateful to Bank Mellat of Iran, specially the directorate, Dr. A. Divandari for his spiritual support and Dr. A. Goli for his advices during the previous years.

I would like to express my gratitude to all the classmates of GFM 2005, especially to Faisal Karim, Changok Lim and Peter Madzudzo, for their encouragement, friendship and all the pleasure that we have shared.

I also appreciate all Iranians of the ITC, specially Dr. A. Abkar, who was the first originator of the JIK(JKIP) program.

Last but not least, my special thanks go to my Iranian classmates, which I learned many things from them during one and a half year study.

Ara Toomanian
Enschede, The Netherlands
March, 2007

Contents

List of Figures

Chapter 1

Beginning of Research

1.1 Motivation

Site selection for banks and other businesses in large cities is important for business enterprise development. The reason is the large investment required for setting up new branches. Previous experience in Iran shows that much money is wasted because of lack of knowledge and strategy for finding new sites. In recent years, business managers have become interested in optimizing their branch locations due to increasing costs. In addition, organizational strategies require the presence of branches all over the region at a sufficiently high density. Therefore, both the number of branches and their distribution are very important.

Several parameters should be taken into account for optimal site selection of branches. These parameters may have spatial, social, technical and economical characteristics, and they can be quantitative or qualitative.

The main goal of this research is to develop a method for discovery of different spatial and non-spatial parameters that have an effect on site selection for bank branches. These parameters may obviously be generalized to other business enterprises, such as insurance companies and fast food restaurants.

In particular, we aim to create a general (spatial) method to help responsible managers, in allocating optimal new branches based on the identified parameters. We are not putting emphasis on site selection as such. What we do here is predicting the efficiency of new branches from existing branch efficiencies regarding to the spatial association rules derived from a number of parameters *before* we do site selection.

A next step is to generalize the model to all business branches, based on specific input and output for any enterprise. On the other hand, there are many important and affecting factors for measuring the efficiency of business enterprises, e.g., profit, income and turnover. We will try to cross-check the derived rules with different dimensions as well as different items mentioned above.

There is another efficiency measurement in this research based on work done using Data Envelopment Analysis in Iran [9, 20]. This concept uses several inputs having different weights based on the opinions of experts and a mathematical model and compares the efficiency of each business branch with the best one. The output is a normalized relative efficiency of branches.

In this research project, we try to establish a relationship between efficiency and spatial aspects by derivation of spatial association rules using methods such as 'A priori' algorithm. For example, if the population size of a region is between 1000 and 2000(or in an average class) and the percentage of commercial land use is more than 30%(or in a high class), then the efficiency of the branch is expected to be more than(0.5) or (in a high class). The rule mentioned above is a vague idea and other templates might be used in future.

1.2 Research identification

The general objective of the research is deriving and testing spatial rules for predicting the efficiency of business branches in different scales such as city region / subregion and branch scales.

1.3 Research questions

This research project tries to answer the following questions:

- What theory of spatial association rules is best applied for the problem statement?
- Which parameter has a better result in the output (spatial association rules) for optimization? (total turnover, efficiency).
- What are the spatial and non-spatial parameters that both positively and negatively influence the target parameters? Besides, which of them are frequently used in the association rules?
- What are the spatial rules derived from identified parameters in different scales?
- How is the spatial association rule implemented for the case of a bank?

1.4 Innovation aimed at

First, we aim at a model for predicting efficiency of new business branches using spatial data. Second, the 'If then' rules generated from the model are another important target result. Actually, the combination of the mathematical model (Operation Research) and algorithms for spatial association rules, such as a priori, is another part of the innovation which we aim to achieve. Third, we use a priori algorithm in such a way that there are classifications inside the regular item set.

1.5 Structure of thesis

This research project includes six chapters. In this first chapter, a general overview about the topic is discussed and the research questions are explained.

Second chapter is a brief description of the basic concepts and literature review for different techniques of data mining and specially association rules. In the third chapter, research area and related data is being discussed. The fourth chapter gives a detailed description of research steps and the methodology in order to achieve the aimed objectives. Chapter five deals with derived rules and results explained in the methodology and the whole research process. Finally, the sixth chapter contains a general discussion and conclusion about the results and also recommendations about the research topics which can be useful for the next generation MSc. students interested in these kinds of subjects.

Chapter 2

Fundamental concepts

This research is a data-centric and database oriented approach. First, *Spatial Data Mining* will be studied in general and various kinds of knowledge extraction methods will be evaluated for this research. Second, the mathematical based model(DEA) as a measurements for efficiency, will be the subject of study. The third dimension of the study is about other related knowledge.

2.1 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 chalenges 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 [12].

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 [13].

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 [12].

Data mining is a general concept and includes various types of algorithms and methods. In this research, although the main issue is the usage of association rule and classification methods but we will provide a brief explanation of some other algorithms.

2.2 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. In this research, these methods are categorized as follows:

- Classification
- Estimation
- Prediction
- Clustering
- Association

The descriptions in this area are extracted from [19].

2.2.1 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', 'Middle', '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. In our example, a new object can be a young male secretary who might be fitted in a low income class. In addition, there are different types of classification methods discussed in the literature which we will explain a bit further. The contents are extracted from [7, 18].

- *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.

In this research project, we will use one of these classification techniques to find better associations between spatial objects. We will try to make an optimal choice of method. In addition, we will aim to always have three meaningful classes for the parameter: 'Max', 'Average', 'Min'.

2.2.2 Estimation

With estimation, one wants 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 $x, t, \ldots$ are the variables, and Y is the total budget estimation.

2.2.3 Prediction

In general, 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 can not 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.

2.2.4 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', 'middle' 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 nice 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. The application of this method will be useful for us, as we are trying to identify different classes.

2.2.5 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* [19].

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*, which will be discussed in Section 2.4.

To have a better understanding of this method, here is an example: In a supermarket with 1000 customers on a special day (Saturday), 500 bought bread and of those, 400 bought butter and jam also. Thus, the association rule would be generated as 'if a customer buys bread then s/he will buy butter'. In this example, the above mentioned parameters of support and confidence are calculated in the Equations: 2.1 and 2.2.

$$\text{support} = \frac{\text{customers bought bread and butter}}{\text{total customers}} = \frac{400}{1000} = 40\% \tag{2.1}$$

$$\text{confidence} = \frac{\text{customers bought bread and butter}}{\text{costumers bought bread}} \frac{400}{500} = 80\% \tag{2.2}$$

As we are going to use this method, these concepts will be discussed in detail in Section 2.4.

2.3 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 [29].

In data mining terminology, three main definitions are considered. An *item* corresponds to attribute-value pair, which in this research 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 [1, 28]:

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 $X \cap Y = \emptyset$.

2.4 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 D that contain X also contain Y. The rule has *support s* in the transaction set D if $s\%$ of the transactions in D contain $X \cup Y$. In other words, probabilistic Formulas 2.3 and 2.4 help in understanding support and confidence.

$$\text{support} = P(X \cap Y) = \frac{\#\text{of transactions containing both X and Y}}{\#\text{of transactions}} \tag{2.3}$$

$$\text{confidence} = \frac{P(X \cap Y)}{P(X)} = \frac{\#\text{of transactions containing both X and Y}}{\#\text{of transactions containing X}} \tag{2.4}$$

In the literature, another way to validate the generated rule is described, and we discuss it in Section 2.8.2.

2.5 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. In the supermarket example mentioned in section 2.2.5, that means that if $\{bread, butter, jam\}$ is frequent, then any subset such as $\{bread, butter\}$ is also frequent. In addition, regarding to the a priori pruning principle: If there is an item set which is infrequent, its superset should not be generated/tested. In the same example, creating a super set of $\{bread, butter, cheese\}$ is not allowed as cheese is an infrequent item.

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 construct a collection of frequent item sets.

2.5.1 The A priori algorithm implementation

Implementation of the algorithm is another important issue in the research area. Using a priori implementation pseudocode, 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 pseudocode represents general implementation method for A priori algorithm [1].

$L_1 \rightarrow$ find frequent 1-item sets(D)
For k in (1, $L_k \neq \emptyset$, k++):
    $C_{k+1} \rightarrow$ candidates for frequent item set generated from (L_k with min-sup)
    $L_k \rightarrow L_1 \times L_{k-1}$
    For each transaction t in D:
    increment the count of all candidates in C_{k+1} if it occurs in t
    $L_{k+1} \rightarrow$ candidates in C_{k+1} with min-sup
return $\cup_k L_k$

Important details of a priori algorithm

As discussed in Section 2.5, the purpose of this algorithm, is frequent item set generation. It has a subprocess 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}.

2.5.2 The a priori algorithm with constraints

In the literature, there is an additional useful concept called a priori+ which is used in cases when one would like to have a restriction on the rules generated.

In other words, the desirable rules are containing a specific object. For such situations, in the frequent item set, tuples which do not contain that object are deleted and the rest is maintained.

In our research example supermarket, the manager is interested in rules about butter. To obtain butter included rules, first we run the a priori algorithm to generate a frequent item set, and then find the biggest L_k that contains butter. In this way, we derive rules where butter is included in the antecedent or consequent [23].

2.5.3 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.

In the example mentioned in Section 2.2.5, $S = \{bread, jam, butter\}$. In this case proper subsets of S are $\{bread, jam\}$, $\{bread, butter\}$, $\{jam, butter\}$, $\{bread\}$, $\{jam\}$, $\{butter\}$. To generate a suitable association rule example, assume $ss = \{bread, jam\}$ then $(s - ss) = \{butter\}$. In this case, $R : \{bread, jam\} \Rightarrow \{butter\}$. The support is the proportion of transactions in which both $\{bread, jam\}$ and $\{butter\}$ occur, which is 400 or $\frac{400}{1000} = 40\%$. Confidence of the rule is 80% due to the fact that the number of customers who bought both bread and jam is equal to the number of customers buying butter. That means, $confidence = \frac{400}{500} = 80\%$.

2.5.4 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 [24].

To reduce the number of rules, 'Lift' and 'Leverage' are two metrics that have been studied in the literature.

Lift

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

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

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 [24]. The lift in this research example equals 2 as the confidence is 100% and support is 50%. In some sources, generated rules ranking method is the same but instead of dividing the two measures, multiplication of support and confidence is calculated [19].

Leverage

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

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

Rules with higher leverage are more interesting than others. In our case, the value for leverage is computed as:

$$Leverage(A \Rightarrow C) = 0.5 - (0.5 \times 0.5) = 0.25$$

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 can not use them during the frequent item set calculation process.

2.6 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 [26].
Spatial data takes the form of *Vectors* or *Rasters*. 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, rasters 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 rasters are good in calculations of and with neighbor pixels. In this research, we work with vector data to represent the topological concepts in the research output and be capable of working with attribute data.

2.7 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 [11].

2.7.1 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' [15]. Spatial topological relationships have a basic role in spatial analysis. In this section we describe some of these concepts, which will be used in this research:

- *Contains / Inside:* These types of relationships happen when a spatial object is completely covering the other. These concept 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.

There are other types of relationships between spatial objects, of which we do not discuss the details [6].

2.7.2 Spatial association rule

A spatial association rule is a rule in the form of $A \Rightarrow B$, where A and B are a set of predicates, some of which are spatial [17]. 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:

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

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 [16]. These concepts were discussed in Section 2.4. In spatial databases, certain topological relationships hold at all times [10]. They can be viewed as spatial association rules with 100% confidence. For example, the containment relationship expressed in Section 2.5 is one such association rule:

$$contains(X, Y) \wedge contains(Y, Z) \Rightarrow contains(X, Z). \tag{2.5}$$

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. For example, one may find that 92% of cities in British Columbia (BC) that are adjacent to water, are close to the U.S.A., as shown in Section 2.6, which associates predicates: *is_a, in* and *adjacent_to* with spatial predicate *close_to* [16].

$$is\_a(x, city) \wedge in(x, bc) \wedge adjacent\_to(x, water) \Rightarrow close\_to(x, USA) \quad (2.6)$$

2.8 DEA

Conceptually, DEA (Data Envelopment Analysis) 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" [27]. By the way, 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 [25]. DEA is a mathematical programming model applied to observation data, which provides a new method of obtaining empirical estimates of extremal 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 [9, 20].

2.8.1 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 can not 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.

2.8.2 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' [8]. 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 [8]. Although this method shows better results, due to the fact that it needs many additional processes in data gathering such as preparing a questionnaire, we will not use this method in this research.

2.9 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 [3]. 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 [4].

2.10 Summary

The concepts discussed in the literature review contained three different major disciplines. First, Data Mining and the different methods and specially association rule method as the main and essential concept in this research. In addition, the relation with the spatial data and the spatial association rules were described briefly. Second, DEA as a kind of mathematical model was covered and the application in measuring the efficiency of any financial institute. Finally, other related fields of science were discussed briefly such as spatial economics. In conclusion, the topics discussed came from different branches of

science but this research tries to combine them. In addition, the term 'DEA-based Spatial association rule mining' will be the innovation of our research in the scientific point of view. In Chapter 3, we discuss the study area and data used as parameter domain and indicate further information about hierarchy between parameters. In addition, we illustrate early data preprocessing related to some network category parameter.

Chapter 3

Research territory

3.1 General information

The study area decided on this research project is the city of Tehran, capital of Iran. The geographical location of Tehran is about $35\,°45'$ North and $51\,°30'$ East. According to the last census data, the population of Tehran is approximately 12 million.

Tehran has 22 municipality zones of which 21 will be analyzed here. One zone is located in the suburbs and most of its area is covered by parks and sport complex, see Figure 3.1. For each municipality zone, there exist subregions,

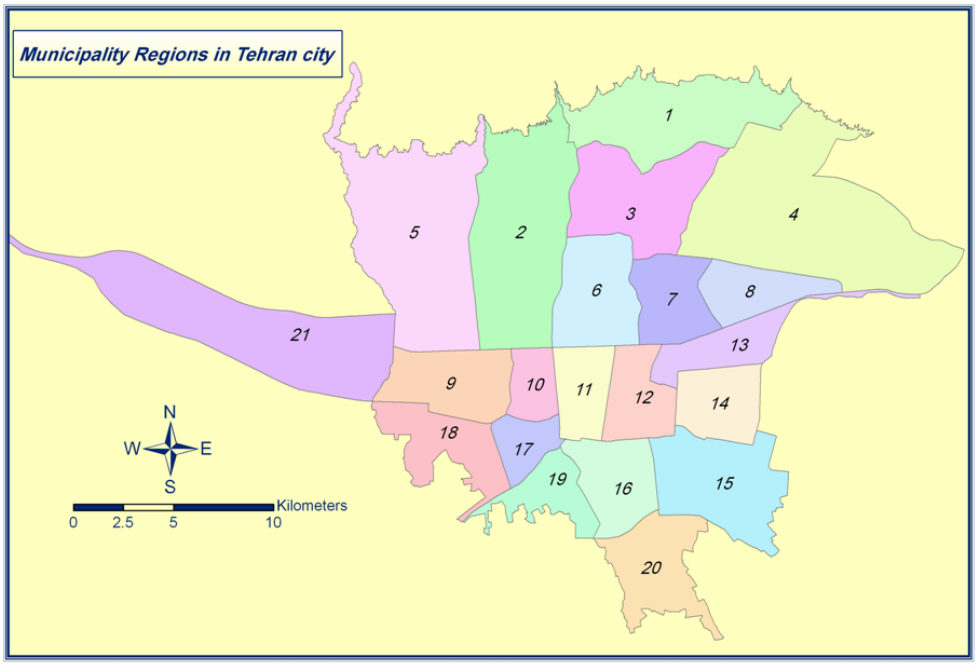

Figure 3.1: City region distribution in the study area

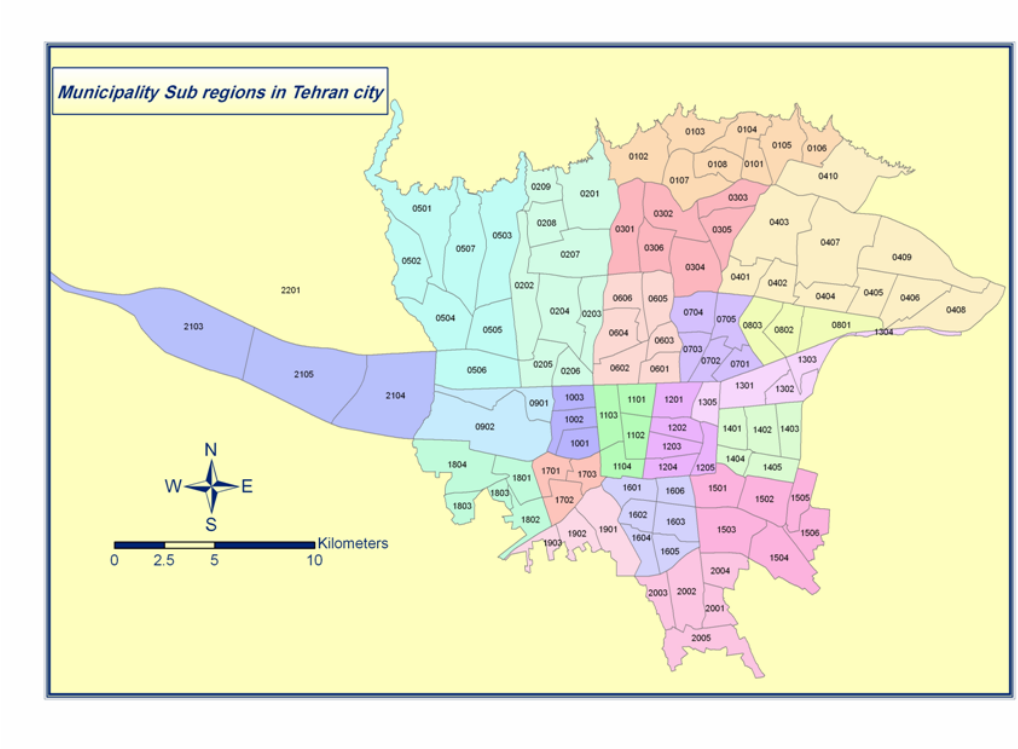

Figure 3.2: City subregion distribution in the study area

which differ in size and number, see Figure 3.2. As mentioned in the first chapter, an application of this research model is in the banking industry. That means, we aim to find relationships between bank branches as a case study and other local spatial parameters. 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 [21, 14, 22, 2, 5]. Essentially, five categories are discussed in most research articles: population, competitors, access, land use, and income. The structure of this research 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, land price for the same region is used in our analysis.

3.2 Bank data

3.2.1 Bank Mellat

The main source data for banks in this research, is one of the governmental banks in Iran called 'Bank Mellat'. It has 303 branches inside 21 municipality regions. Using Global Positioning System (GPS) technology, all branch coordinates have been 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, first, a relative comparison between branches called 'DEA based efficiency' is 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 2.8. We take into account two additional more sensible measurements to clarify abstract efficiency concept. The total turnover and number of customers 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 help to evaluate and validate the results.

3.2.2 Competitors

Two banks were selected as competitors for this category. Both were selected from governmental banks, which more or less have the same number of branches. The parameter used in this category was competitor location. Non-spatial data for the competitors has been 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.

3.3 Population data

Various types of census data used in this research are provided by the National Population and Housing Census Data Bureau. Different attributes in census data consist of total population, number of employed, unemployed and literate people in each zone. Furthermore, additional attributes have been derived from main data such as ratio of discussed values per total population.

3.4 Land use data

For each scale in the research, there are three different types of land use data, collected from municipality tax department. 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.

3.5 Trade area

In addition to the land use data, the number of retail shops and big malls are other parameters used in this research for the three scales. This factor is extracted from national cartographic center of Iran, based on 1:2,000 maps.

3.6 Network data

Network data is an important parameter for communication between business branches. In region and subregion scales, it is classified in three different categories as 'freeways', 'main roads' and 'secondary roads'. 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.

3.6.1 Shortest path in network data

From the network parameter perspective, in the branch scale, we use additional network calculations for two point parameters. We use the shortest path concept to find the minimum distance between bank branch points and urban facilities such as police stations, in the role of a network parameter. In this research, we use the buffer of 5000 meter to find nearest police station to each branch. The calculated distance is used for classification in three classes as 'high', 'average' and 'low' distance, see Figure 3.3. In addition, we use the dis-

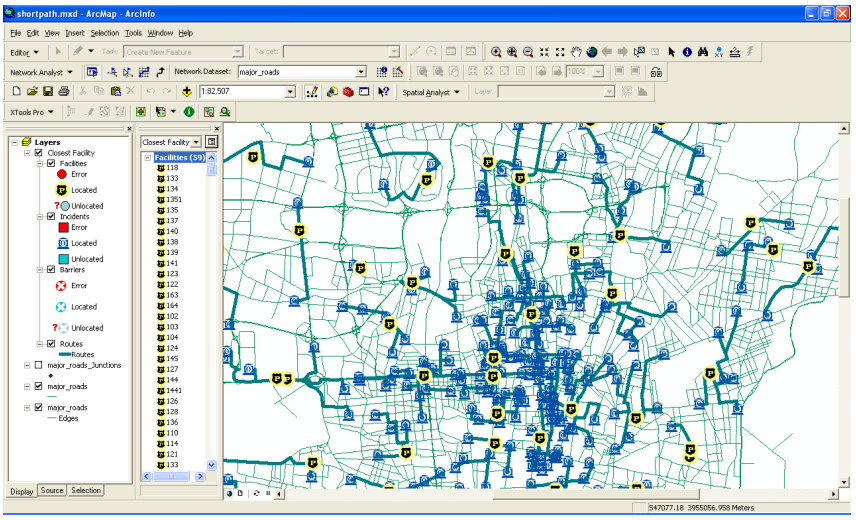

Figure 3.3: Bank Mellat branch distance from police stations

tance between Bank Mellat as an internal bank with two competitors. With a buffer of 5000 meter, minimum distance between them is measured, and also classified like the previous one. The method used for classification in this category, is also natural breaks. The result of shortest path competitions is shown in Figure 3.4.

3.7 Land price data

An average land price is calculated for each region and subregion and also for the branch scale. As discussed before, land price is used instead of the income

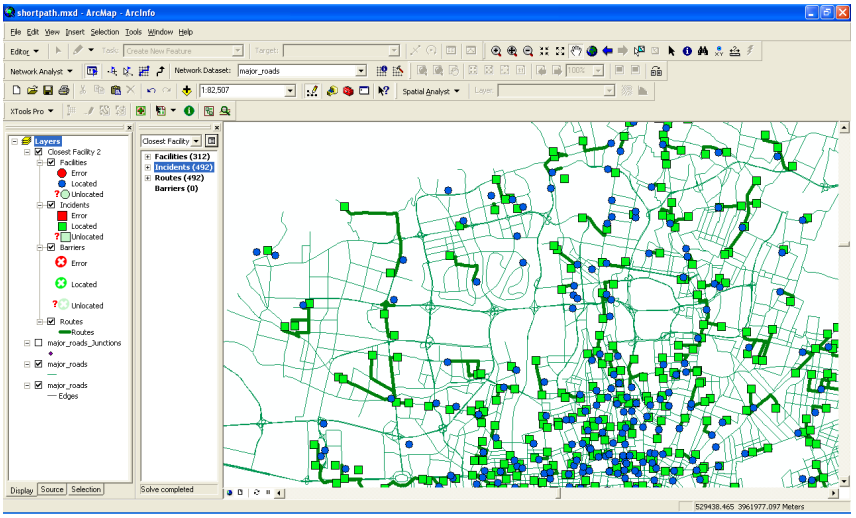

Figure 3.4: Bank Mellat branch distance from the competitors

parameter to fill the gap in the economic category. In this parameter, the value is classified in three classes as 'high price', 'average price', and 'low price' .

3.8 Parameter hierarchy

In this research, we aim to analyze the association between parameters in city region / subregion and branch scale. Different parameters used in each mentioned scales are shown in a hierarchy, see Figure 3.5.

In some parameters such as population, the factors are the same due to the data gathering limitation, and may cause some problems. An ideal design is to define a border for each branch and search for the research parameters inside a buffer for the branch scale, but as far as we do not have such a database in those data (like the US TIGER database) we have to trust the existing data. In the appendix, a brief description of the parameters used in the process is represented.

3.9 Summary

In the research domain, different city scales explained in this chapter and also their contribution was discussed. Afterwards, the parameter categories were briefly described. In some domains, early additional calculations needed for the next stages were clarified. Later on, the parameter hierarchy in all scales gave us a general overview of the spatial factors.

In chapter 4, we use a step-by-step description for the methodology. Besides, an a priori like algorithm of which the basic part was discussed in previous chapters, will be explained briefly. In addition, an alternative way for decreasing the parameter domain will be described. Then two different scenarios related to the output of the algorithm will be discussed.

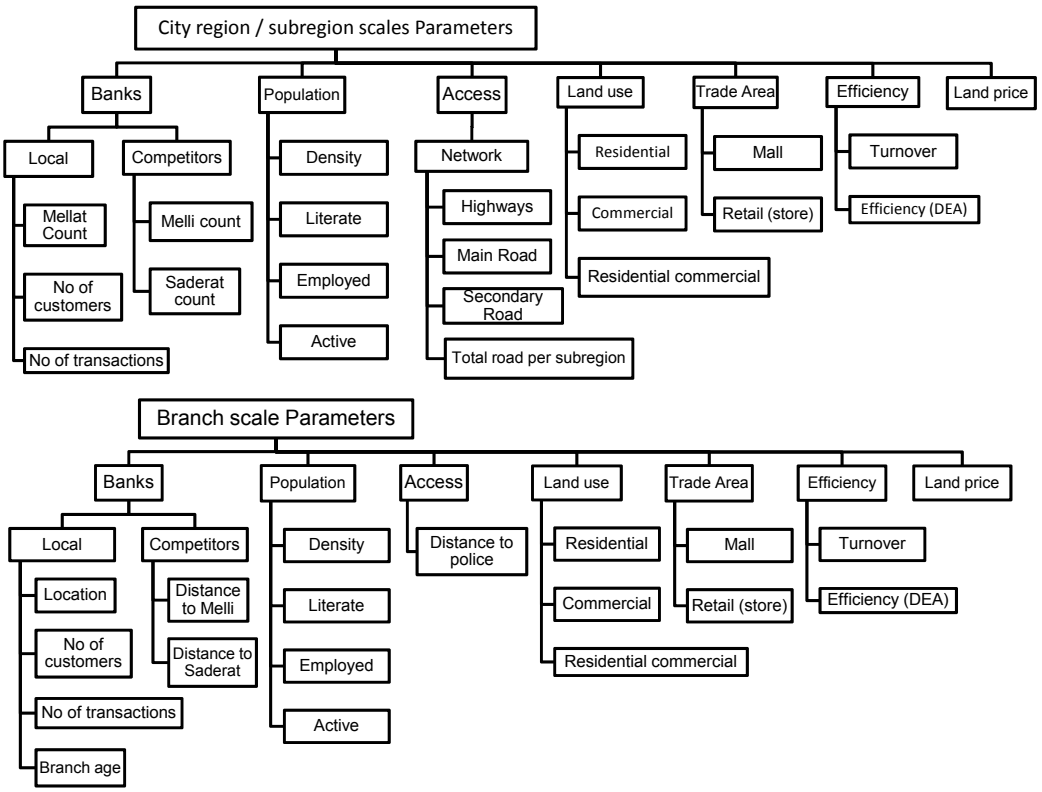

Figure 3.5: Parameters used in different scales

Chapter 4

Research: Step-by-step

4.1 Introduction

In this research project we try to combine non-spatial measurement with a set of different spatial parameters mentioned in previous chapter, to find associations between spatial parameters to predict efficiency in different spatial scales. In addition, we also aim to find similarities and differences of derived spatial rules in both city regions / subregions and branch scales.

4.2 Data preprocessing

As discussed in previous chapters, and regardless of output structure, a number of additional variables derive from existing parameters. For a better understanding of the concept, here is an example. In the population dimension, size of the literate population is one of the parameters which by itself is an important variable but if we calculate the ratio of literate population to total population, the result is even more useful and understandable. In this way, there is an early step in the method called data preprocessing, to prepare essential inputs for the main method.

4.2.1 Role of efficiency parameters in the association rule

At the early stages, we tried to find association rules in the whole study area but because the support and confidence of the rules generated were very low, we aimed to find the potential sites in which the efficiency is 'high', 'average', or 'low'. The reason is the determination of approximate areas where the above mentioned measure is nearly in the same range. To do so, a local polynomial interpolation method is used for this process.

Primary results shows that, in north and west parts of the study area, there is a spatial pattern for high efficiency, while the center of the study area contains average efficient branches and there is a low efficient spatial pattern in the southern part of the city, see Figure 4.1.

In addition, there is another measurement included in the research to compare the abstract efficiency with another sensible quantity. The same process

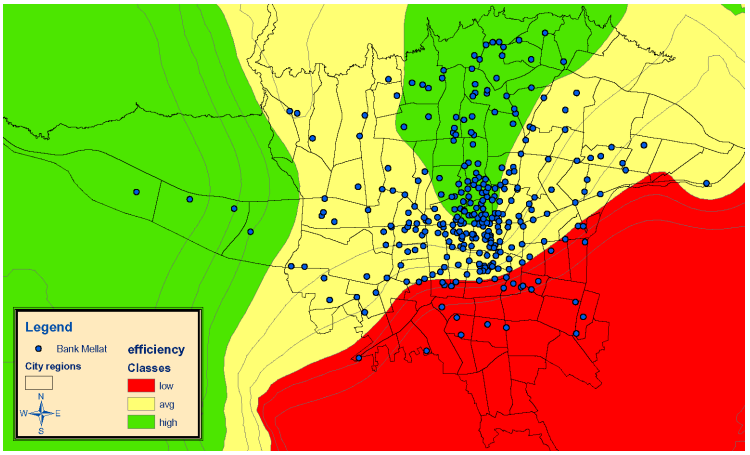

Figure 4.1: Spatial patterns of efficiency in the study area

is done for the turnover of each branch, which originates from the amount of deposit of each account. The result of comparison shows substantial spatial similarity between two measurements, see Figure 4.2 .

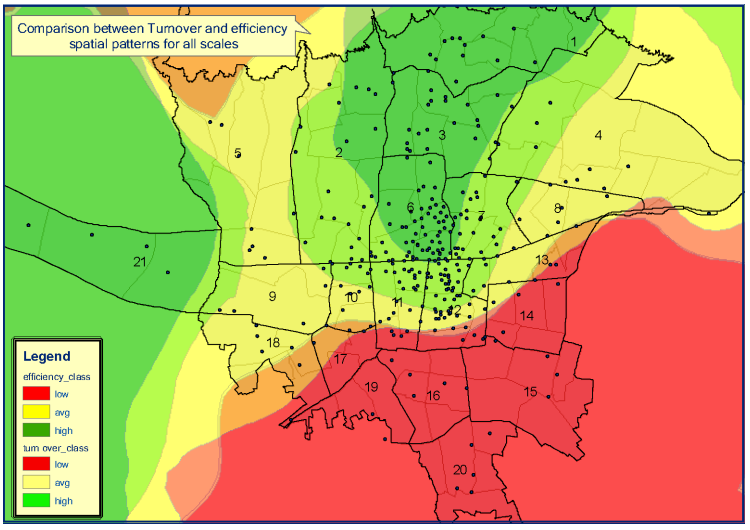

Figure 4.2: Comparison between two efficiency measurements

4.2.2 Spatial parameters used in the research

Spatial data used in this research originates from different social, economical, business and also infrastructure resources. At the city region and subregion 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 subregion are calculated in the process and also, polygon layers are merged at the subregion level.

4.2.3 Layer transformation to PostgreSQL database

Since we are interested in spatial phenomena playing a role in efficiency determination, we are using a spatial database as an integrated media to store our data.

4.2.4 Spatial queries for data preprocessing

In the previous sections, aggregation of different layers was discussed briefly. In PostGIS, we use topological relations such as 'contains' and 'inside' not only to overlay geographical layers in the database, but also to join and calculate spatial parameters without any primary or foreign key relationship. In addition, for those layers that cannot completely be overlayed due to existing errors such as topological and multiple data sources, we use polygon centroids for proper spatial joins. The following query is a simple example of the method applied here.

```
SELECT R.regions_id, SUM(L.x1) as residential,
 SUM(L.areax1) as area_residential, SUM(L.x2) as commercial,
 SUM(L.areax2) as area_commercial, SUM(L.x3) as res_commercial,
 SUM(L.areax3) as area_res_commercial
INTO regionlevel_landuse
 FROM municipality_regions as R, landuse as L
WHERE L.the_geom && R.the_geom
 AND contains(R.the_geom, centroid(L.the_geom))
GROUP BY R.regions_id
ORDER BY R.regions_id
```

In this query, which is an example for such cases, we have a spatial join between city region and land use layer and using the centroid of land use layer, calculate the number of different types of land use in a city region.

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, we should classify every parameter. As discussed in Section 2.2.1, we select two different classification methods to evaluate which is more useful for research data. The result of comparison shows that, the natural breaks method is more useful than quantile. This is because of the nature of the efficiency data. As it is represented in the figure 4.3, the histogram of the efficiency data is in such a way that, most of the branch efficiency data is less than 0.4, so it is completely a right-skewed curve.

After data preprocessing step, all the parameters have a three way classification: 'Max','Average' and 'Min'.

4.4 Main table implementation

The main table is the result of data preprocessing. It holds all association parameters for each scale. In other words, it is a table with variable classification

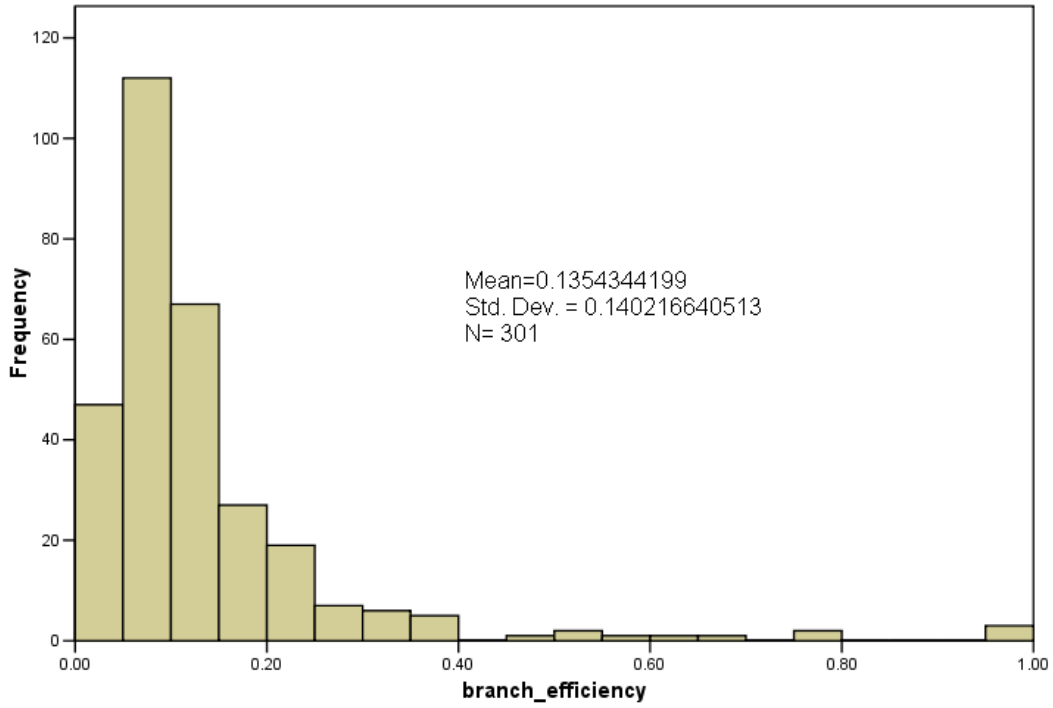

Figure 4.3: DEA based efficiency parameter histogram

at the parameter transaction level. In this table, all the characteristics of a region / subregion or branch scale are described in a single tuple, see Figure 4.4. For additional description about main table, we represent a part of a tuple. In the city subregion scale, we have a subregion with the characteristics: Total branches of Bank Melli is average, total branches of Bank Saderat is maximum, total length of secondary roads is average, total length of main roads is average, total length of freeways is average, literate population size is minimum, number of retail stores is maximum and etc. In mentioned tuple there are thirty two different spatial and non-spatial parameters which explain the general characteristics of that subregion.

In the usual a priori method, there is no classification preprocess used for the objects. In other words, all data is boolean. For example, in a supermarket, the customer buys butter or does not buy butter so in the object transaction tuple for that customer there exists $\{\text{butter}(b)\}$ or does not exist an item as $\{(\neg b)\}$. Here in this research, as we classify a wide range of parameters in three classes, there is always a characteristics of that parameter in the main table as 'parameter (min)', 'parameter (avg)' or 'parameter (max)'. In addition, we keep track of the parameter names in the main table. This extra calculation is not necessary in the normal a priori algorithm. The reason to do so is the triple classification method used in research.

After the main table generation, we have to choose tuples corresponding to the efficiency / turnover classes mentioned in Section 4.2.1. For instance, in the branch scale, using a spatial join, all branches inside the high efficient potential areas are selected and separately stored in a different table for subsequent

id_number [PK] text	melli text	saderat text	len_3 text	len_4 text	len_5 text	meelat text	max_rto text	avg_rto text	min_rto text
0101	melli (min)	saderat (avg)	len_3 (min)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (max)	avg_rto (min)	min_rto (min)
0102	melli (min)	saderat (min)	len_3 (max)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (max)	avg_rto (min)	min_rto (min)
0209	melli (min)	saderat (min)	len_3 (min)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (min)	avg_rto (min)	min_rto (max)
0301	melli (avg)	saderat (avg)	len_3 (avg)	len_4 (avg)	len_5 (avg)	mellat (avg)	max_rto (avg)	avg_rto (avg)	min_rto (avg)
0302	melli (avg)	saderat (min)	len_3 (avg)	len_4 (avg)	len_5 (max)	mellat (avg)	max_rto (max)	avg_rto (avg)	min_rto (min)
0303	melli (min)	saderat (min)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (min)	avg_rto (min)	min_rto (min)
0304	melli (avg)	saderat (avg)	len_3 (max)	len_4 (avg)	len_5 (max)	mellat (avg)	max_rto (max)	avg_rto (min)	min_rto (min)
0305	melli (avg)	saderat (max)	len_3 (max)	len_4 (min)	len_5 (min)	mellat (avg)	max_rto (avg)	avg_rto (avg)	min_rto (min)
0306	melli (min)	saderat (min)	len_3 (avg)	len_4 (max)	len_5 (avg)	mellat (avg)	max_rto (avg)	avg_rto (avg)	min_rto (avg)
0401	melli (min)	saderat (avg)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (avg)	avg_rto (min)	min_rto (avg)
0402	melli (min)	saderat (min)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (min)	avg_rto (avg)	min_rto (avg)
0403	melli (min)	saderat (min)	len_3 (max)	len_4 (min)	len_5 (max)	mellat (min)	max_rto (min)	avg_rto (max)	min_rto (min)
0404	melli (min)	saderat (avg)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (avg)	max_rto (min)	avg_rto (max)	min_rto (avg)
0405	melli (min)	saderat (min)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (min)	avg_rto (max)	min_rto (min)
0406	melli (min)	saderat (min)	len_3 (avg)	len_4 (min)	len_5 (min)	mellat (min)	max_rto (min)	avg_rto (min)	min_rto (min)

Figure 4.4: The main table representation

steps. The same procedure is also implemented for other scales and classes of efficiency measures. At the end of this stage we will have six different tables with specifically those tuples that have highest probability for each efficiency classes. Then, as we need an array data structure of the a priori like algorithm implementation, we should convert parameters into into an array data structure. This stage can be 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
```

At this stage, we have obtained the input for the a priori like algorithm implementation. Basic part of the a priori algorithm was discussed in previous chapters 2.5, but the a priori like algorithm used in this research has minor changes which will be explained in Section 4.5.

4.5 The a priori like algorithm implementation

In general, we use the infrastructure of the a priori algorithm discussed in Chapter 2.5 for this research but the idea of our algorithm, which has been implemented with the Python programming language, has differences from the regular algorithm and uses extra computations. These dissimilarities are described in three cases as follows:

1. We should check the situations where the algorithm finds association between different classes in parameters and then reject them. For instance, suppose we have a parameter called land price in the subregion scale. If we find two subsets of parameters being frequent, such as ‘avg price’ and ‘high price’ in one frequent item set, this is meaningless. The reason is that in the final rule we will have an area with both average and high price which is not logically correct. To solve this problem, we need a kind of indexing system, to keep track of the parameter name, in which whenever we use a subset of any parameter from the list, other subsets will never be used in the frequent item set selection.

2. Regarding the a priori algorithm, to generate L_k (k-frequent item set), we should join the L_{k-1} set with L_1 as discussed in the Chapter 2.5.1. As the array data structure used of this research is in text data type and string array manipulation is rather complex and difficult, we have to use extra computations for the process such as additional functions for sorting the elements which is not necessary for a digit variable.

3. Another difference between the regular a priori algorithm and our technique is that we have to sort the output arrays to delete any repetitions. For example, if the second frequent item set contains two arrays such as {a,b}, {b,c} and the first frequent item set include {c}, {a} then the result of the third frequent item set will have duplicates ({a,b,c}, {b,c,a}) and one must be omitted.

At the end of the algorithm implementation most frequent item set is generated. The source code of the program is represented in the appendix.

4.6 Parameter reduction

The idea of parameter reduction is to try find in each parameter category a subset of parameters with meaningful similarities. The reason to reduce parameter is that we have more than thirty parameters in each scale. This amount of spatial and non-spatial factors cause a long process time for the algorithm. In such cases, if we can find a parameter that can replace from others, no only the number of parameters will decreased but also the process time will be reduced.

In this stage we survey the possibility of parameter reduction both in conceptual and practical point of view.

Conceptually, we can define a sequence between the parameters of a category. In the population class, for instance, in this data set there are four parameters. Regardless of measure classification such as 'high', 'average' and 'low', we put more emphasis on the nature of the parameters. This means, we only test to find a sequence between population size, employed population, literate population, and unemployed population. In a common region, one would expect such a sequence, see Figure 4.5.

In this diagram, two scenarios are represented. First, if the population size increases then the literate population size will be also increase. In such a case, increasing the employed population size will decrease the unemployed population size. For the second scenario, the first part is the same but from the literate population size there is a direct connection to the unemployed population size. Needless to say, such scenarios come from the rules derived from a data set and might be different in other data collections. In this case, we can ignore the unemployed population size to reduce the number of parameters.

In a practical point of view, we may deduce a causal sequence between population category parameters, using the rules derived from the a priori algorithm.

A total scheme is first to find the validation measures for each rule, then in each step, select one with the largest leverage and process until the last frequent item. For example, in the population data we can find such a causal

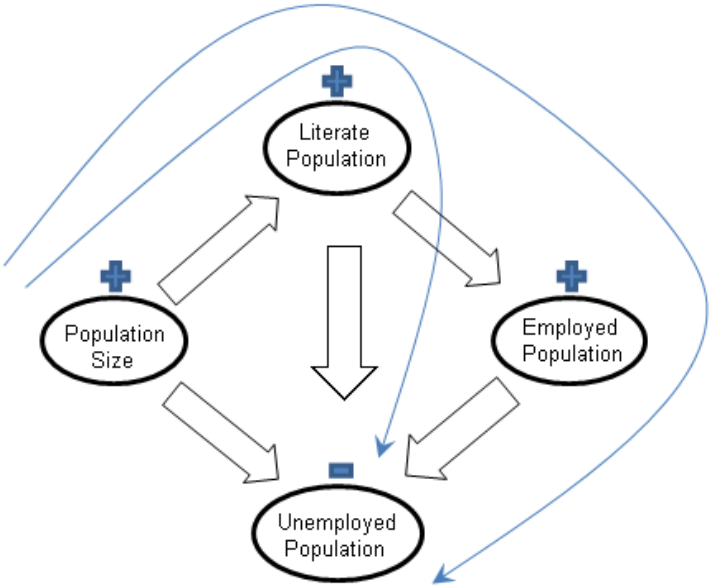

Figure 4.5: Conceptual sequent in the population category

sequence via following steps. The population category has four subparameters: Population size as (pop), employed population as (pop_work),literate population as (pop_litr), unemployed population as (pop_nowork). At the first step, they form the input for the a priori algorithm. The result of the algorithm, which is the frequent item set, is shown in Figure 4.6. In this table, arrparval contains the most frequent item set for each step of the a priori algorithm, cnt field, is the number of repetition of each set, and col column, contains the indexing system for parameters used in the algorithm.

As it is represented in the table, there are three different 'Third-item sets' as the result of the a priori algorithm. The next step is to find all possible association rules and calculate the support and confidence, lift and leverage for them. The result of this process is as follows: Frequent item set: *Pop(max), Pop_litr(max), Pop_work(max)*

$$Pop(max) \Rightarrow Pop\_litr(max)$$

Support=34%, Confidence=67%, Lift=1.93, Leverage=0.3492

$$Pop\_litr(max) \Rightarrow Pop\_work(max)$$

Support=34%, Confidence=100%, Lift=1.35, Leverage=0.0919

$$Pop(max) \Rightarrow Pop\_work(max)$$

Support=51%, Confidence=100%, Lift=1.35, Leverage=0.1357

If we start from $Pop(max)$, there are two alternative rules, where the one with higher lift and leverage is selected, and after that there exists only one rule so the only candidate is being picked for the sequence, see Figure 4.7.

arrparval [PK] text[]	cnt double precis	col smallint[]
{pop(max)}	161	{2}
{pop(max),pop_litr(max)}	109	{1,5}
{pop(max),pop_litr(max),pop_work(max)}	109	{1,3,5}
{pop(max),pop_nowork(min)}	110	{1,7}
{pop(max),pop_nowork(min),pop_work(max)}	110	{1,3,7}
{pop(max),pop_work(max)}	161	{1,3}
{pop(min)}	151	{2}
{pop(min),pop_litr(min)}	151	{1,5}
{pop(min),pop_litr(min),pop_nowork(min)}	128	{1,5,7}
{pop(min),pop_nowork(min)}	128	{1,7}
{pop_litr(max)}	109	{6}
{pop_litr(max),pop_work(max)}	109	{5,3}
{pop_litr(min)}	203	{6}
{pop_litr(min),pop_nowork(min)}	167	{5,7}
{pop_litr(min),pop_work(max)}	121	{5,3}
{pop_nowork(min)}	238	{8}
{pop_nowork(min),pop_work(max)}	163	{7,3}
{pop_work(max)}	230	{4}

Figure 4.6: Frequent item set table for the population category

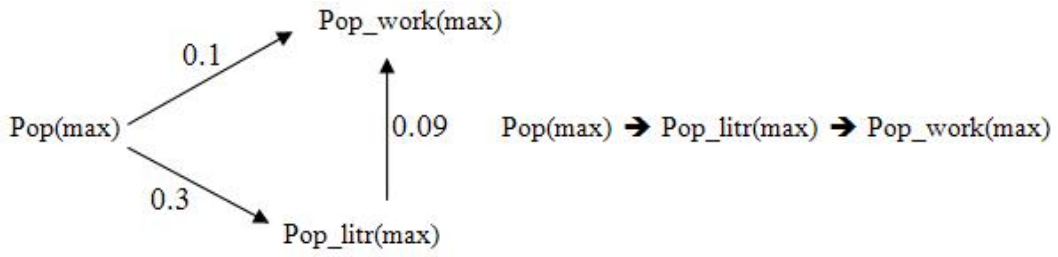

Figure 4.7: Sequence for the frequent item set in population parameter

Frequent item set: *Pop(min), Pop_litr(min), pop_nowork(min)*

$$Pop(min) \Rightarrow Pop\_litr(min)$$

Support=48%, Confidence=100%, Lift= 1.53, Leverage=0.1690

$$Pop\_litr(min) \Rightarrow Pop\_nowork(min)$$

Support=54%, Confidence=82%, Lift= 1.07, Leverage=0.039

$$Pop(min) \Rightarrow Pop\_nowork(min)$$

Support=41%, Confidence=85%, Lift= 1.11, Leverage=0.0411

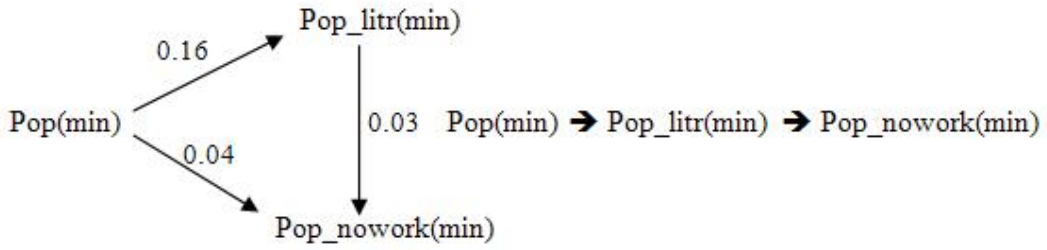

Figure 4.8: Sequence for the frequent item set in population factor

Again, if we start from $Pop(min)$, there are two alternative rules, and the one with higher lift and leverage is selected. After that there exists only one rule so the only candidate is being picked for the sequence, see Figure 4.8.

Frequent item set: *Pop(max), Pop_nowork(min), Pop_work(max)*

$$Pop(max) \Rightarrow Pop\_nowork(min)$$

Support=36%, Confidence=68%, Lift= 1.156, Leverage=0.0411

$$Pop\_work(max) \Rightarrow Pop_nowork(min)$$

Support=52%, Confidence=97%, Lift= 1.26, Leverage=0.0398

$$Pop(max) \Rightarrow Pop\_work(min)$$

Support=52%, Confidence=70%, Lift= 1.27, Leverage=0.1840

Also, if we start from $Pop(max)$, there are two alternative rules, and the one with higher lift and leverage is selected. After that there exists only one rule so the only candidate is being picked for the sequence, see Figure 4.9.

To summarize this reduction part, in two cases we found a causality between Pop and Pop_nowork we use the Pop parameter instead of Pop_nowork. This method can be used for other categories.

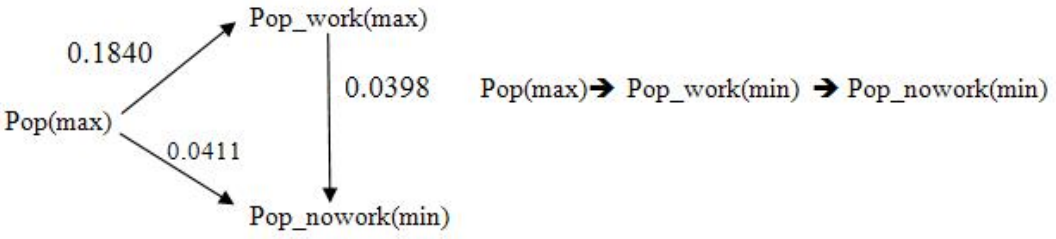

Figure 4.9: Sequence for the frequent item set in population domain

4.7 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 than 75 item set 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 so good to generate frequent item set and the process was continued with the remaining scales.

4.8 Association rule generation

As discussed in the literature review of Section 2.5.3, for association rule generation we need to produce all subsets of the frequent item set.

The outputs of the a priori like algorithm in each scale, are candidates for spatial association rule implementation. We also discussed about the constraints existing for some certain parameters in Chapter 2.5.2. At this stage, we have a constraint related to the consequent in the association rules derived from the frequent item set.

As we are looking only for the efficiency parameters in the obtained rules, we have to retain the tuples including the most frequent item sets of the efficiency / turnover parameters, and delete the rest. In this way, the most frequent item set related to the efficiency / turnover will be obtained. The rules for each scale are listed and discussed in the next chapter.

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 subregions / 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, we can measure the efficiency of the branches inside those subregions or rely on the number of the high, avg 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 Summary

In this chapter, we discussed about all processes that derive association rules. In an early step, using a local polynomial interpolation, three major spatial po-

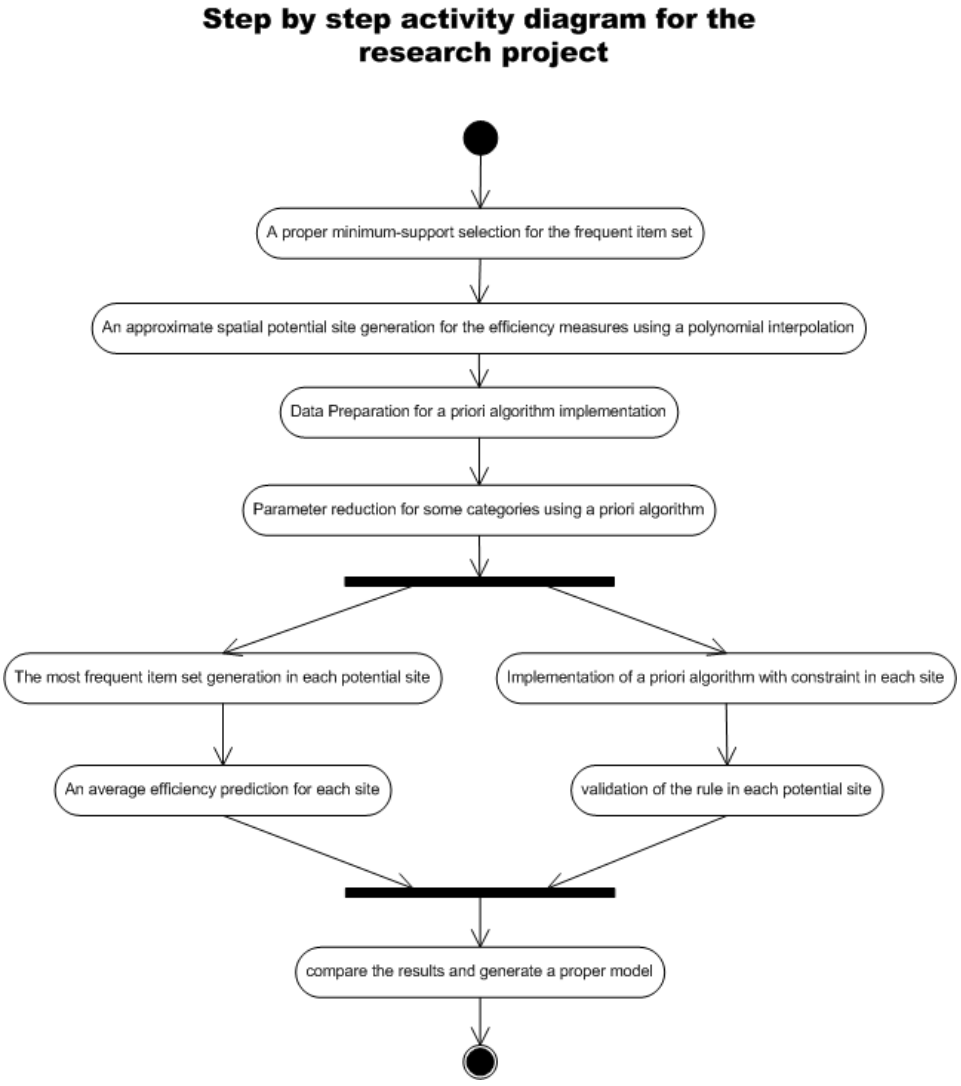

Figure 4.10: Step by step of the research

tential sites were selected and for each class the process is done for generating association rules. Afterwards, using an a priori like algorithm the frequent item set is generated. Then, due to the fact that we are looking for an efficiency measurement, we select those frequent item sets that include an efficiency parameter. The next stage contains rule generation with the subsets of the frequent item set and calculation of validation concepts. In addition, all parameters existing in the frequent item sets, regardless of their contents, are compared with all table tuples and an average amount of efficiency is calculated.

In this chapter, we also suggest a method to decrease the number of parameters in a category based on the a priori algorithm. The activity diagram in UML notation for this research, describes essential steps for the whole research. In this diagram, both scenarios have been represented in a high level, See figure 4.10.

In the next chapter, we present the results of two scenarios in different

scales and also two efficiency measures.

Chapter 5

Research results

Introduction

The content of this chapter is omitted due to the limitations of the source data and local regulations.

For more Information please contact toomanian14996@itc.nl.

Chapter 6

Discussion and recommendations

6.1 Introduction

In this research project, we aimed to find spatial patterns for a non-spatial measure. At first stages, we tried to find a spatial pattern in the whole study area but due to the nature of the data set, the result was not so good. In addition, support and confidence of the rules were low. Using an approximate spatial pattern for the efficiency measures, the rule validity concepts increased. We successfully found all frequent item sets for each spatial class of potential site to predict the efficiency measure.

To have a rule including an efficiency measure, regarding the association rule method, the efficiency measure must be involved in the frequent item set so we have to put a constraint on the frequent item sets to complete to process and calculate the validity parameters.

In this chapter, we first discuss about the derived and generated results and then after a brief conclusion, we recommend topics for additional future work.

In the discussion section, we will discuss the results from three different perspectives. First, from a parameter point of view, we evaluate the results and also try to find their weak points. Secondly, multiple scales used in this research are discussed, to compare differences in small and large scales. Thirdly, we indicate the usability of the method and provide a general overview of results.

6.1.1 Parameter perspective

- In most of the derived rules, competitor location is highly related to the efficiency. This means, due to the market sharing concepts, when the density of competitor is low, we should expect a low efficiency site. The above mentioned situation, in this data set, happens almost in the subregion scale data.

- Low frequent parameters are omitted at the first step of algorithm due to the a priori algorithm. They almost include all positive and high class elements.

- Certainly, the lack of positive parameters will cause a negative weight on the results which is a special case in our research and will be different for other data sets. In addition, we cannot expect to have these parameters in the result of prediction.
- In this research, we had a limitation of detailed data related to the local economical parameters such as average income of people, total investment of each region.
- Due to the a priori property, all subsets of the frequent item set could be the antecedents of the rules so that it is possible to derive rules with less parameters included.
- For such data sets in which the parameters are not normally distributed, this method can derive rules where the efficiency density is high and the outliers do not exist. In this research project, as it was represented in Figure 4.3, most of the samples are less than 0.4 so we expect low class and negative rules from the data set.

6.1.2 Different scales perspective

- At the city region scale, frequent item set was not generated even with some limited elements because using a large number of parameters for a limited regions, it is difficult to find a proper association between items based on a minimum support.
- At the branch scale, distance to the competitor is minimum due to the fact that distribution of competitors are similar to the local branches in this data set.
- Comparing the two scales, the branch scale has rules with higher lift and leverage for the high efficiency parameters but in turnover measure, the subregion scale has higher lift and leverage because we could easily aggregate the measure from the small scale data.
- At both city subregion and branch scales, almost the same parameters include in the rules. This means, the same set of parameters exist in the most frequent item set and just the network concept changed in these scales. At the branch scale, due to the lack of data, the distance_to_ competitor and police station were two network parameter proxies but in the city subregion scale this concept changed to the total length of freeway, main road and secondary road.

6.1.3 Results and methods perspective

- With this method, we can predict efficiency of branches in which the parameters are involved with the rules. In other words, due to the nature of 'If then' rules, we will be able to predict efficiency measure when the antecedent happens.

- For this data set, only for the 'city subregion low efficiency', we could find two types of consequents such as low and average efficiency due to the aggregation of data **??**. In other words, for that situation, low efficiency and average efficiency parameters were frequently happen.

- Quantile classification method, to classify the efficiency parameters, produced equal tuples in all classes. In a high range class, tuples with an approximate same value were in low class which was unforseen. In this case, rules derived from the quantile classification method are not promising.

- Run time process for the low efficient class was longer than the other classes not only due to the more calculations but also more frequent elements in the results.

- For all scales, most of the low efficient measures generate rules in which the leverage is 1. In general, lift values greater than 1 indicate that the consequent is more frequent in transactions containing the antecedent than in transactions that do not. In our research, due to the fact that low measure is occurring in all tuples, so there is no tuple remaining to compare with others.

6.2 Conclusion

In this research project, we aimed to find a prediction model for the efficiency based on the spatial association rules. Using such a model, the managers of a business branch, are able to make better decisions for to 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, we 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 involve 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 a kind of negative rules in all the classes. In this way, for such data collections with a right skewed curve, we suggest areas which are not efficient. The comparison of 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.

During the research, we find 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.

6.3 Recommendation

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 blocks. For branch scale we 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, we suggest to find alternative methods for data sets which have not a normal distribution and have Poisson distribution.
- We also recommend to use an alternative method of spatial association rule based on parameter weight, where in the small scales finds a total weight of each parameter and due to that weight for the larger scales try to implement the same rule with giving additional weights to some parameters.
- This research was based on a snapshot of time for both the efficiency measurements and spatial parameters. We strongly recommend to add time dimension to the research and think about a temporal spatial association rule. In this case, we can also combine the temporal spatial association rule with a visualization method such as space-time-cube to detect the spatial temporal changes in a city.

Appendix A

Code of the a priori like algorithm

```
import pg

import time

min_sup=11

# Selection of a minimum support for the frequent item sets

time.clock()

db=pg.DB(dbname='postgres',host='localhost',user=
'postgres',passwd='########', port=5431)

# truncating tables are needed for making sure previous

# values are deleted

db.query('truncate x') db.query("truncate f1") db.query("truncate
frequentn") db.query("truncate frequentm")

paramlist=['melli','saderat','len_3','len_4','len_5',
'meelat','max_rto','avg_rto','min_rto','res_a','com_a',
'res_com_a','res_rto','com_rto','res_com_rt','pop','pop_lit'
,'pop_work','litr_rto','work_rto','over10_class','road_tot'
,'rod_tot_rt','mall_s','store','trans','customer',\ 'price_rang']

# paramlist is a list of the field names for each measure

l=len(paramlist)
k=2 for u in paramlist:
    nice= db.query('SELECT %s,%s
    FROM sb_region_original_income_low
    GROUP BY %s '%(u,k,u)).getresult()
    k=k+1
    db.inserttable('x',nice)
```

```
# x is a kind of index table to keep track of fields as

# there is a triple system in the field values

for u in paramlist:
    db.query("DELETE from x where parval='{%s}'"%(u))

# omitting any repetitions k=1 for u in paramlist:
    nice2=db.query('SELECT %s,count(*),%s
    FROM sb_region_original_income_low
    GROUP BY %s having count (*) >%s
        '%(u,k,u,min_sup)).getresult()
    k=k+1
    db.inserttable('f1',nice2)
    db.inserttable('frequentn',nice2)
# first frequent item set and also first subset of n-th

#frequent item set generation

step1=db.query("SELECT array[f1.arrparval,f2.arrparval],
count(a.*),array[f1.col,f2.col] FROM f1, f1 as f2,
sb_region_array_income_low as a WHERE f1.arrparval < f2.arrparval
and f1.col <> f2.col and array[ f1.arrparval ,f2.arrparval] <@
a.item GROUP BY array[f1.arrparval , f2.arrparval],
array[f1.col,f2.col] having count(*) >%s
ORDER BY array [f1.arrparval,f2.arrparval]"%(min_sup)).getresult()
db.inserttable('frequentn',step1)

# joining the first frequent item set to have the second frequent
item set

for n in range(3,l):
    c=db.query("INSERT INTO frequentm SELECT
    (fn.arrparval || array[f1.arrparval] ),count(*),
    (f1.col || fn.col) FROM sb_region_array_income_low
    as a, f1, frequentn as fn WHERE a.item[f1.col]=
    f1.arrparval AND fn.arrparval <@ a.item and
    array_upper(array[fn.arrparval],2)=%s and NOT(f1.col =
    ANY(fn.col)) GROUP BY fn.arrparval,f1.arrparval,
    (f1.col || fn.col)having count(*) > %s ORDER BY
    (fn.arrparval || array[f1.arrparval])"%(n-1,min_sup))

    # next steps for filling in the n-th frequent item set

    # a swap table to remove repetitions from the n-th frequent item set
    db.query("INSERT INTO frequentn SELECT distinct
    (array_sort(arrparval))
    as arrparval,cnt,array_sort(col) FROM frequentm as
    fm WHERE array_upper(array[fm.arrparval],2)=%s"%(n))
    db.query("truncate frequentm")
cl= time.clock()
print cl
```

Appendix B

Parameter description in different scales

B.1 The city Subregion scale parameter description

Field name	*Description*
id_number	A unique id for each subregion
melli	Bank Melli density
saderat	Bank Saderat density
len_3	Secondary road length in a subregion
len_4	Main road length in a subregion
len_5	Freeway length in a subregion
meelat	Bank Mellat density
max_rto	Maximum efficiency ratio per branches
avg_rto	Average efficiency ratio per branches
min_rto	Minimum efficiency ratio per branches
res_a	Residential land use area in a subregion
com_a	Commercial land use area in a subregion
res_com_a	Residential commercial land use area in subregion
res_rto	Residential land use area ratio per total land use
com_rto	Commercial land use area ratio per total land use
res_com_rt	Residential commercial land use area ratio per total land use
pop	Population size in subregion
pop_lit	Literate population size in a subregion
pop_work	Employed population size in a subregion
litr_rto	Literate population size ratio per total population in a subregion
work_rto	Employed population size ratio per total population in a subregion
over10_class	Active population size or the population more than 10 years old
road_tot	Total road length in subregion
rod_tot_rt	Total road length ratio per area in subregion
mall_s	Number of malls in a subregion
store	Number of retail stores in a subregion
trans	Total transaction in a subregion
customer	Number of customers in a subregion
price_rang	Land price in a subregion

B.2 The branch scale parameter description

Field name	*Description*
branch_id	Branch code
income	Total turnover
”day”	DEA based efficiency measure
total_transaction	Total transaction of a branch
”location”	The location of a branch (North, Center, South)
age	Branch age
mellatcust	Total customers of a branch
price_rang	Approximate land price of a branch
close_to_saderat	Distance between a local branch and the nearest Bank Saderat branch
close_to_melli	Distance between a local branch and the nearest Bank Melli branch
mlat_to_po	Distance between a local branch and the nearest police station
population	Population size beside a branch
density	Population density beside a branch
pop_litrac	Literate population size beside a branch
pop_worker	Employed population size beside a branch
litrac_ratio	Literate population ratio per population beside a branch
work_ratio	Employed population ratio per population beside a branch
over10_class	Active population size or the population size more than 10 years old
residential_area	Residential land use area beside a branch
commercial_area	Commercial land use area beside a branch
res_com_area	Residential commercial land use area beside a branch
residential_ratio	Residential land use area ratio per total land use beside a branch
commercial_ratio	Commercial land use area ratio per total land use beside a branch
res_com_ratio	Residential commercial land use area ratio per total land use
tot_store	Number of malls beside a branch
store	Number of retail stores beside a branch

Bibliography

[1] R. Agrawal, T. Imieliski, and A. Swami. 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, 1993. ACM Press.

[2] N. Al-Hanbali. Building a geospatial database and GIS data-model integration for banking: ATM site location. International Society for Photogrammetry and Remote Sensing (ISPRS), September 2003.

[3] L. Anselin. What is special about spatial data? alternative perspectives on spatial data analysis. in spatial statistics: Past, present, and future. *Ann Arbor*, pages 63–77, 1990.

[4] L. Anselin and A. Getis. Spatial statistical analysis and geographic information systems. *Annals of Regional Science*, 26:19–33, 1992.

[5] M. Birkin and G. Clarke. GIS, geodemographics, and spatial modeling in the U.K. financial service industry. *Journal of Housing Research*, 9(1), 1998.

[6] P. A. Burrough. *Principles of Geographical Information Systems for Land Resources Assessment*, chapter 2. Clarendon, 1986.

[7] J. Campbell. *Map Use and Analysis*. McGraw-Hill, fourth edition, 2001.

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

[9] A. Divandari, G. R. Jahanshahloo, and F. Hosseinzadeh Lotfi. 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, 2006.

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

[11] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann series in Data Management Systems. Morgan Kaufmann, first edition, August 2000.

[12] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd edition, March 2006.

[13] D. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, Cambridge, 2001.

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

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

[16] K. Koperski. *A progressive refinement approch to spatial data mining*. PhD thesis, Simon Fraser University, April 1999.

[17] K. Koperski and J. Han. 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, 1995. Springer-Verlag.

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

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

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

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

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

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

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

[25] R. Ramanathran and R. Ramanathan. *An Introduction to Data Envelopment Analysis*. SAGE Publications, 2003.

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

[27] B. Scales. Data envelopement analysis: a technique for measuring the efficiency of govenment service delivery. Technical report, Australian Government productivity Commission, 1997.

[28] R. Srikant and R. Agrawal, editors. *Mining generalized association rules with item constraints*, volume KDD 97. Third International Conference on Knowledge Discovery and Data Mining, 1997.

[29] N. Ye. *The Handbook of Data Mining*. Lawrence Erlbaum, Mahwah, New Jersey, 2003.