Aalto University
School of Engineering

SoftGIS Data Mining and Analysis: A Case Study of Urban Impression in Helsinki

Master's Thesis

Aalto University
School of Engineering,
Department of Real Estate, Planning and Geoinformatics,

Espoo, 27 March 2014

Kamyar Hasanzadeh
Master of Science in Geoinformatics

Supervisor: Professor Kirsi Virrantaus
Instructor: D.Sc.(Tech) Jussi Nikander
D.Sc.(Tech) Paula Ahonen- Rainio

Aalto University
School of Engineering

Aalto University, P.O. BOX 11000, 00076 AALTO
www.aalto.fi

Abstract of master's thesis

Author	Kamyar Hasanzadeh
Title of thesis	SoftGIS Data Mining and Analysis: A Case Study of Urban Impression in Helsinki
Degree programme	Geomatics
Major/minor	Geoinformatics
Code of professorship	Maa-123
Thesis supervisor	Kirsi Virrantaus
Thesis advisor(s)	Jussi Nikander, Paula Ahonen-Rainio
Date	27.03.2014
Number of pages	48+2
Language	English

Abstract

In recent years there has been considerable breakthrough in acquisition of qualitative georeferenced data. SoftGIS is one of the most prominent attempts in this context that is capable of providing useful data that has applications in different disciplines. However, similar to any other large spatial dataset, the SoftGIS data requires a set of spatial analysis and data mining techniques in order to yield the desired information and to be considered as a reliable source of knowledge.

This thesis propounded a four stage knowledge discovery process in which several exploratory, visual and analytical spatial techniques were proposed. Moreover, the proposed techniques were applied and implemented in a case study of urban impression in Helsinki metropolitan area. The proposed techniques take advantage of an appropriate data quantification approach and aim to facilitate the knowledge discovery process of spatio-qualitative data and contribute to the revelation of the desired information.

Due to the distinct characteristics of this type of data, partially caused by its acquisition procedure and partially by its qualitative nature, certain considerations needed to be taken into account. These considerations could not be accomplished without studying the data and exploring its specific characteristics. The striking existence of cognitive uncertainty in SoftGIS data led to the application of fuzzy logic techniques in this thesis. The results indicate that using fuzzy techniques is a promising approach towards mitigating the negative effects of the aforesaid uncertainty in SoftGIS datasets.

Furthermore, this thesis widened its domain of knowledge discovery to a less explicit realm of information through employing spatial data mining. This resulted in discovery of interesting associations between the SoftGIS data and the neighboring building types.

Keywords SoftGIS, knowledge discovery, Spatio-qualitative data mining, association rules, fuzzy, clustering validation, spatial analysis, visualization

Acknowledgment

I would like to express the deepest appreciation to my supervisor Professor Kirsi Virrantaus, for her guidance during my studies at Aalto University and for giving me the opportunity to work in the department of Geoinformatics.

My special gratitude goes to my instructor Dr. Jussi Nikander for his brilliant guidance and the useful comments, remarks and engagement through this Master's thesis. Many thanks also go to my kind and brilliant teacher and instructor Dr. Paula Ahonen-Rainio for her encouragement, useful comments, and for all she taught me during my Master's studies.

Furthermore I would also like to acknowledge with much appreciation and respect the crucial role of Professor Vesa Niskanen for his kind encouragements and suggestions during my master thesis work.

I would also like to thank Jaakko Rantala for his friendliness and contribution of valuable comments to my thesis as well as all my other colleagues at the department of Geoinformatics for an inimitable and friendly working atmosphere.

Finally I would like to express my sincere thanks to my dearest parents and other members of my family, for all their love and immense support.

And the last but not the least, I would like to thank the Mapita Ltd. group for providing the SoftGIS dataset, without which this thesis could not be accomplished.

Espoo, March 2014
Kamyar Hasanzadeh

Table of Contents

Abbreviations

SDM	Spatial data mining
ARM	Association rule mining
CH	Calinski-Harabasz index
DB	Davies–Bouldin index
RMSE	Root mean square error
fcm	Fuzzy c-means clustering
EDA	Exploratory data analysis
WAV	Weighted average visualization

List of Figures

List of Tables

1. Introduction

1.1 Background

Studying qualitative data in their geographical context has the potential to reveal useful information in different studies, such as human geography, geology, urban studies and land use planning. Accordingly, there has recently been an increasing interest in applications of spatial technologies, and more specifically GIS, in studying qualitative data (Verd & Porcel, 2012). A geographic information system (GIS) is a collection of hardware, software, and data for exploring and analyzing all types of geographically referenced information. In other words, GIS provides us with wide range of tools and techniques that can help us to gain a better understanding of different phenomena in their geographical contexts.

Recent advances in spatial sciences and computer technology have allowed qualitative GIS to be incorporated in the latest versions of computer-aided qualitative data analysis (Verd & Porcel, 2012). This has contributed to the considerable growth in acquisition of geocoded qualitative data. The Finnish innovation 'SoftGIS', developed by Marketta Kyttä and her team at Aalto University, refers to a collection of internet-based surveys (Figure 1) that allow the 'locality-based' study of human experiences (Kahila & Kyttä, 2009, Kahila & Kyttä, 2006). SoftGIS provides a combination of 'soft' subjective data (qualitative) with 'hard' objective spatial data and is capable of collecting large datasets for the use of urban planners and other professionals interested in the development of more user-friendly physical settings (Kahila & Kyttä, 2009; Rantanen & Kahila, 2009).

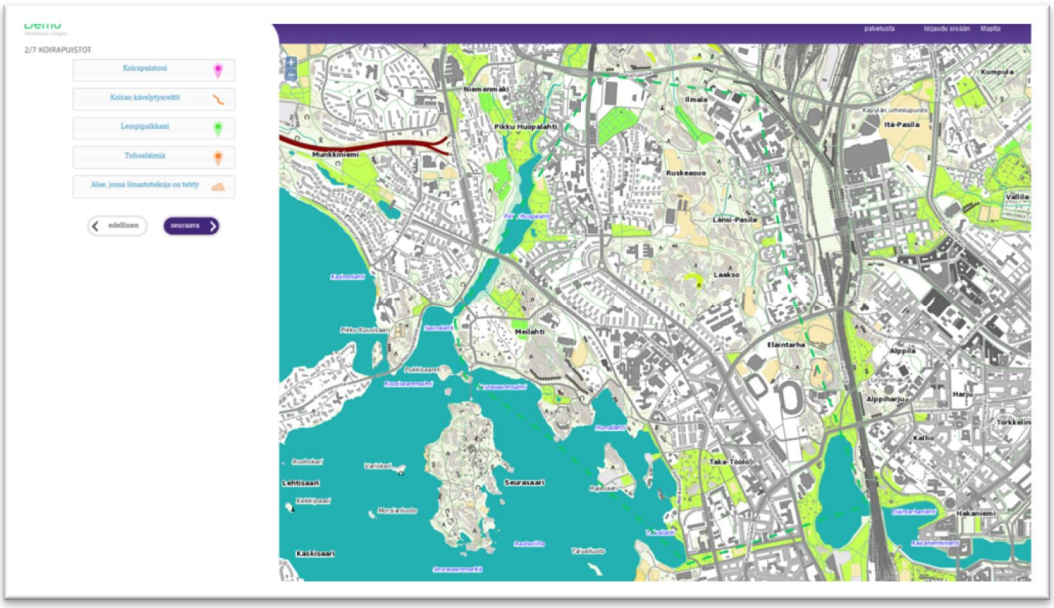

Figure 1. SoftGIS survey panel preview*: http://demo.asiatkartalle.fi/ (accessed on 5.2.2014)*

1.2 Objective and research questions

With the considerable increase in professionals' access to soft geocoded data, the need for a solid analysis methodology has been raised; one that suits the specific characteristics of the SoftGIS data and has the potential to reveal the most useful information from the huge mass of data.

Most of the existing literature regarding SoftGIS, address the SoftGIS data acquisition methodology, and there have not been much work on data analysis and knowledge discovery of this type of data. However, there have been some studies using typical GIS exploratory data analysis (EDA) techniques in order to study SoftGIS data (Kyttä *et al.*, 2013; Kahila, 2008). Although the approaches purposed and implemented in the mentioned literature reveal useful information, they have major limitations that prevent a comprehensive understanding. For instance, some of the methods used in the aforesaid literature capture an overly generalized image of the circumstances. This contradicts the principal characteristic of SoftGIS which is to be 'locality-based' (Kahila & Kyttä, 2009). Moreover, in some cases the visualizations used are not sufficiently informative and can even cause misinterpretation.

This thesis is based on the idea that by knowing specific characteristics of SoftGIS data, existing GIS analysis techniques can be further improved and adapted so that they can reveal more interesting knowledge from the data. This thesis is motivated by this idea and it aims to respond to the following research questions:

"What is specific about SoftGIS data?" Prior to applying any knowledge discovery technique, this thesis begins with an analysis of the data in order to provide a better understanding of the SoftGIS data characteristics which will be used as basis of the further proceedings.

"Can a suitable visualization of the given dataset reveal useful information?" In other words, this thesis aims to examine the best visualization techniques for the given dataset and explore the knowledge that can be visually discovered in the region of study.

"What useful information can be revealed through spatial data mining techniques?" In other words, this thesis aims to explore useful patterns and knowledge that can be discovered using spatial data mining techniques.

1.3 Methods and materials

SoftGIS data reflects a large number of people's perceptions and experiences of spatial locations (Kahila & Kyttä, 2009). Therefore, it can be perceived as a large source of information that needs to be processed and analyzed scientifically in order to generate useful knowledge. This cannot be achieved without a solid understanding of the SoftGIS data specifications. This understanding can be obtained through a wide range of techniques that are known as spatial analysis. Spatial analysis or spatial statistics includes any of the formal techniques which study entities using their geometric or geographic properties (O'sullivan & Unwin, 2003).

Subsequently, this thesis purposes a number of visual, analytical, exploratory and data mining techniques that are tailored to meet the characteristics of SoftGIS data or any other similar qualitative geocoded dataset. Furthermore, the proposed methods are implemented in a case study of urban impression in Helsinki metropolitan area, namely Espoo and Helsinki, and the discoveries are discussed. On whole, this thesis uses a four stage spatial knowledge discovery strategy which is adapted to meet SoftGIS data specifications (Figure 2).

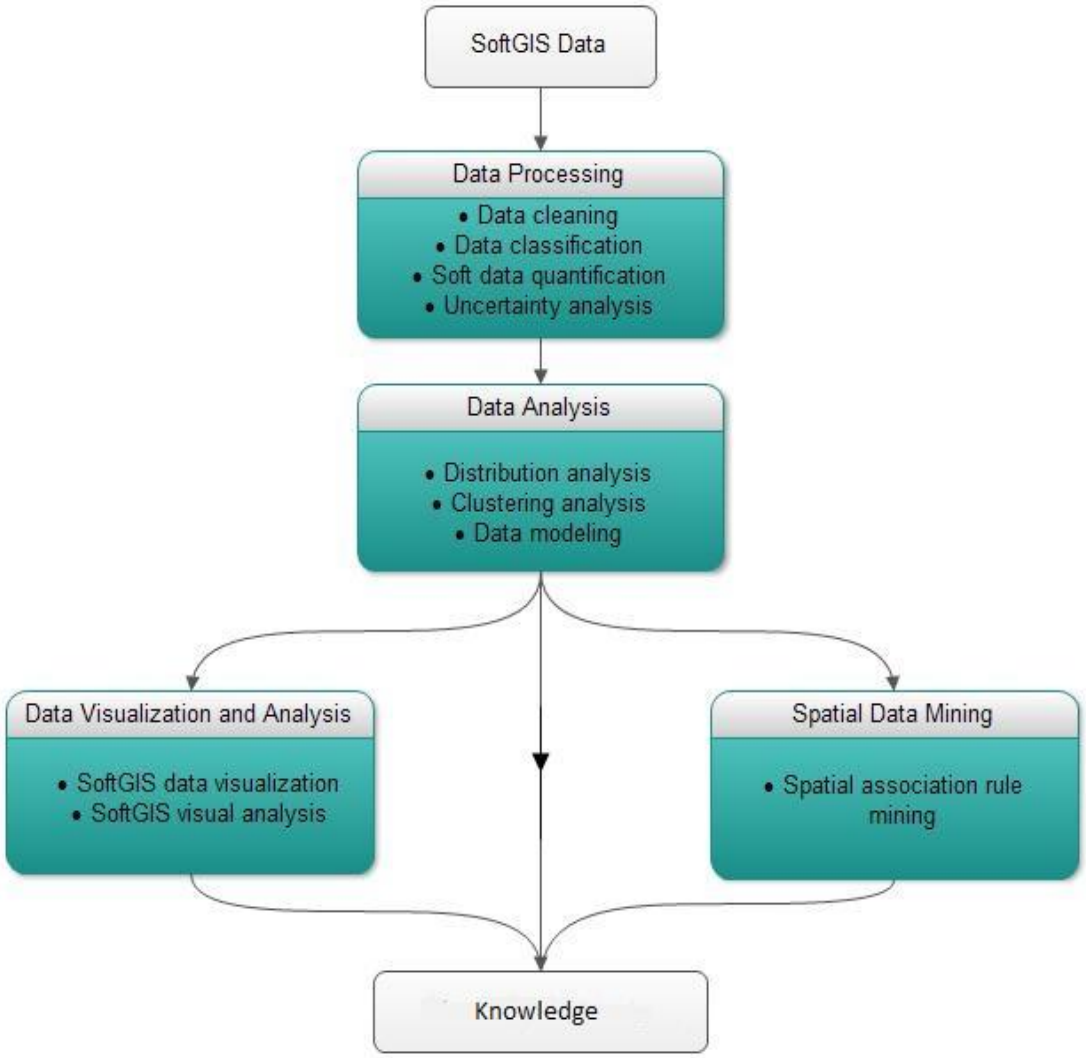

Figure 2. SoftGIS knowledge discovery process

Two datasets are used in this study: Helsinki region SoftGIS data and SeutuCD (building dataset):

- The SoftGIS data was provided by Mapita Ltd. and it included people's recorded urban experiences in Espoo and Helisnki region, Finland. The dataset originally consisted of six classes of records, three classes of positive and three classes of

negative records (atmosphere, appeal, social). For the simplicity of the study the classes were generalized into two major classes of positive and negative records.

- All municipalities in Finland are obligated to collect register data on their population, buildings and land use plans. Since 1997 the Helsinki Metropolitan Area Council (YTV) is working on the production of a database covering the whole Helsinki Metropolitan area with data from the municipality registers. The outcome is a data package called SeutuCD that includes register data of population, buildings, agencies and enterprises as well as the data related to land use planning and real estate (HSY website).

Studying qualitative data is normally more challenging than working with quantitative datasets. According to this, in disciplines dealing with qualitative values (e.g. social and psychological sciences), it is usually a good practice to quantify the values in order to improve the computational and comparison capacity of the qualitative datasets (Guttman, 1944). The computational techniques widely used for quantitative data are typically inapplicable to qualitative datasets as they contain nominal values. Computationally-enabling the qualitative data facilitates the implementation of a diversity of visual and analytical techniques that help to make the analysis more feasible. Therefore, in this study, in order to overcome the existing computational limitation, a simple quantification is used. Thus, the values +1 and -1 are assigned to the nominal *Positive* and *Negative* quantities respectively.

It should also be noted that in this thesis the following software have been used:

- *ArcGIS 10.*1 for spatial analysis and visualization,
- *Matlab R2013b* for modeling, clustering, and other computational tasks,
- and *IBM SPSS Modeler 15.0* for association rule mining.

1.4 Thesis structure

Following the *Introduction*, the thesis continues with the chapter *Theory background* which is a theoretical reference to all the computational concepts and methods used throughout this thesis. This chapter is intended to provide the reader with a more solid understanding of the processes via covering further theoretical details on these techniques.

Subsequently, the third chapter *Data analysis* explores the specific characteristics of SofGIS data and aims to gain an overall understanding of the data as well as its dominant trends and patterns. This chapter begins with a data distribution analysis using Moran's index and continues to a spatial cluster analysis and a comparison of the results from soft (fuzzy) and hard clustering methods. Moreover, in this chapter, two distinct models, namely fuzzy and polynomial, are fitted to the dataset and compared with an aim of finding a better behavioral description of the given SoftGIS dataset.

Chapter four, *Data visualization,* begins with a discussion of existing visualization methods for representing SoftGIS data and then proceeds to develop the idea by proposing and implementing alternative approaches in order to tackle currently existing limitations.

However, due to the large size of the dataset, the traditional spatial analysis techniques and visualizations are not capable of capturing all existing patterns. Therefore, spatial data

mining is proposed as an appropriate approach for detecting further interesting patterns in the dataset. Spatial data mining (SDM) is a knowledge discovery process which is used to extract implicit interesting information, spatial relations, and any other kind of knowledge that is not explicitly stored in dataset (Koperski *et al.*, 1996; Koperski & Han, 1995; Leung, 2010). Thus, the fifth chapter applies a spatial data mining technique, namely spatial *Association rule mining* or co-location, in order to discover more profound and statistically supported associations and patterns in the given dataset.

2. Theoretical background

In this chapter each computational method and concept used throughout this thesis will be briefly presented. This chapter covers the principal theories on data distribution measures, root mean square error, clustering methods, clustering validation, and association rule mining.

The aim of this chapter is to provide the gist of the used concepts and methods in order to facilitate the interpretation of the results and to contribute to the better understanding of the reader.

2.1 Data distribution measures

This section covers further theoretical details on the measures of distribution used in section 3.1. More specifically, this section elaborates on Moran's I and correlation coefficient *r*.

2.1.1 Moran's I

In statistics, Moran's Index (also known as Moran's I) is a measure of spatial autocorrelation that was developed by Patrick Alfred Pierce Moran (1950). Mathematically speaking, Moran's I is defined as:

$$I = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j} z_i z_j}{S_0 \sum_{i=1}^{n} z_i^2}$$

Where Z_i is the deviation of an attribute for feature I from its mean ($x_i - \bar{X}$), $\omega_{i,j}$ is the spatial weight between feature i and j, n is the total number of features, and S_0 is the aggregate of all the spatial weights:

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} \omega_{i,j}$$

Moran's I values range from −1 to +1, each indicating perfect dispersion and perfect correlation respectively while a zero value indicates a random spatial pattern (Moran, 1950). Moreover, negative values indicate negative spatial autocorrelation and positive values indicate positive spatial autocorrelation. A positive autocorrelation implies that the closer the two objects are the more similar they would be. On the other hand, a negative autocorrelation implies that the closer the two objects are the less similar they would be. It is also good to note that in spatial datasets, positive autocorrelation is more common than negative autocorrelation.

2.1.2 Correlation coefficient, *r*

The quantity *r*, called the linear correlation coefficient, also known as Pearson's product-moment coefficient, measures the strength of the linear association between two variables (Dowdy & Wearden, 1983). *r* is obtained by dividing the covariance of the two variables by the product of their standard deviations.

For a series of *n* measurements of X and Y written as x_i and y_i where i = 1, 2, ..., n, *r* can be defined as:

$$r = \frac{1}{n-1}\sum\left(\frac{X-\overline{X}}{S_X}\right)\left(\frac{Y-\overline{Y}}{S_Y}\right)$$

Where $\overline{X}$ and $\overline{Y}$ are the means of X and Y, and S_x and S_y are the standard deviations of X and Y. It should be noted that scales or units of measurement are not a concern while calculating *r*. That is because the formula for calculating the correlation coefficient standardizes the variables.

The correlation coefficient always takes a value between -1 and 1, with 1 or -1 indicating perfect correlation. A positive correlation indicates a positive association between the variables, while a negative correlation indicates a negative association between the variables. Positive correlation implies that increasing values in one variable correspond to increasing values in the other variable. On the other hand, a negative correlation indicates that increasing values in one variable corresponds to decreasing values in the other variable. Furthermore, a correlation value close to 0 indicates no association between the variables (Dowdy & Wearden, 1983).

2.2 Root mean square error (RMSE)

The Root Mean Square Error (RMSE), which is also often referred to as 'The Root Mean Square Deviation' (RMSD), is frequently used to assess models' quality and it is a measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. RMSE serves to aggregate these individual differences, or so-called residuals, and is mathematically defined as (Hyndman *et al.*, 2006):

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

Where X_{obs} and X_{model} are respectively the observed and modelled values at *i*.

The RMSE values are easy to interpret in a sense that according to the formula above, a larger difference between the observed value and the model value results in a bigger RMSE. In other words, a large RMSE value casts doubt on the validity of the model while

a small RMSE implies its efficiency. Obviously, the notion of larger or smaller is relative and therefore should be assessed based on the original values' range.

2.3 Clustering analysis

As described in Section 3.3 several clustering analyses are performed using four distinct clustering algorithms. The four algorithms used in 3.3, namely SOM, K-means, hierarchical, and fcm, take different approaches towards capturing the clustering characteristics in a dataset. More specifically, *SOM* is a method based on an artificial neural network that attempts to preserve topological properties, *k-means* is a centroid-based method which aims to minimize the within cluster differences, *hierarchical* is connectivity-based and aims to capture the existing patterns within the dataset, and finally fcm, which is similar to k-means but it is based on fuzzy logic and takes advantage of gradual membership of each point to the clusters.

The following sections provide further details on these methods.

2.3.1 Self-organizing map (SOM)

The Self-organizing Map (SOM), first proposed by Teuvo Kohonen, is an artificial neural network that is capable of distinguishing similarity patterns in multidimensional space. However, SOMs are different from other artificial neural networks in the sense that they preserve the topological properties of the input space by using a neighborhood function (Špatenková, 2009). There are several implementations of SOM for clustering purpose.

The result of SOM is generally a good approximation to the input space as it attempts to preserve topology and approximate the data density of the input dataset (Costa, 2010).

SOM is briefly used in this study as one method of identifying clusters in a spatial data (Section 3.3). However, SOM is a very wide topic and discussing it in more details is out of scope of this study. Interested reader is suggested to read more about this topic in (Špatenková, Demšar, and Krisp, 2007; Špatenková, 2009).

2.3.2 K-means clustering

K-means algorithm is one of the most common clustering techniques and there are numerous implementations of it. In general, given the appropriate number of clusters *k*, the algorithm aims to minimize the object function *J* (within-cluster sum of squares) of each cluster (MacQueen, 1967; Tan, Steinbach, and Kumar, 2005):

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| x_i^{(j)} - c_j \right\|^2$$

Where $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centroid c_j.

Consequently, k-means clustering results in *k* clusters where the within cluster deviations are minimized. Nevertheless, k-means typically results in clusters of approximately similar

sizes and depending on the data, it may yield incorrectly cut borders in between of clusters, as the algorithm always assigns an object to the nearest centroid. In other words, the k-means algorithm mostly aims to optimize the cluster centers rather than cluster borders (MacQueen, 1967).

2.3.3 Hierarchical clustering

Hierarchical clustering is a method of cluster analysis that attempts to build a hierarchy of clusters. Hierarchical clustering groups the data over a variety of scales by creating a dendrogram (Manning *et al*., 2008; Johnson, 1967). There are generally two strategies for hierarchical clustering:

- *Agglomerative* or "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- *Divisive* or "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

In this study an agglomerative approach is used.

In general, hierarchical clustering is a connectivity-based method that attempts to capture the existing patterns in the dataset while maintaining the minimum similarity requirement.

2.3.4 Fuzzy c-means clustering

As in fuzzy logic (Zadeh, 1965), in fuzzy clustering, in contrast to the conventional methods where every point completely belongs to just one cluster, each point has a degree of belonging to clusters. Thus, points on the edge of a given cluster may be to a lesser degree in the cluster than points in the center of cluster (Ruspini, 1970).

Any point x has a set of coefficients giving the degree of belonging to the kth cluster $W_k(x)$. With fuzzy c-means, the centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster (Kaymak, 2000; Dunn, 1974):

$$C_k = \frac{\sum_x w_k(x)^m x}{\sum_x w_k(x)^m}$$

The degree of belonging, $W_k(x)$, is related to the inverse distance from x to the cluster center. It also depends on a predefined parameter m that controls how much weight is given to the closest center (Kaymak, 2000; Klawonn, 2008). The fcm algorithm minimizes intra-cluster variance and it typically results in smoother and more realistic cluster boundaries which is a result of gradual membership. However, it has the problem that the minimums are local minimums and the results are highly dependent on the initial choice of weights m (Ahmed *et.al*., 2002).

2.4 Cluster validation

Cluster validation is a term referred to the process of evaluating clustering results. There have been many suggestions for a measure of similarity between two clusterings and they can be used to compare how well different data clustering algorithms perform on a set of data.

There are two groups of clustering validation measures, namely internal and external evaluation. In internal evaluation, a clustering result is evaluated based on the data that was clustered itself (Manning, Raghavan, and Schütze, 2008). On the other hand, in external evaluation, clustering results are evaluated based on some data that was not used for clustering, such as known class labels or external benchmarks (Färber *et al.*, 2010).

In this study as there is no external data available, three of most common internal clustering indices are used to compare the clustering quality in section 3.3. It should be noted that these measures are usually used to get some insight into whether one algorithm performs better than another, and this shall not necessarily imply that one algorithm produces more valid results than another as these measures are usually tied to the type of criterion being considered in assessing the quality of a clustering method and are often biased towards algorithms that use the same cluster model (Manning, Raghavan, and Schütze, 2008; Maulik, 2002).

2.4.1 Calinski-Harabasz

Calinski-Harabasz is an index based on the ANOVA (Analysis of variance between groups) ideology. The Calinski-Harabasz index (CH) is defined as (Maulik, 2002):

$$CH = \frac{SS_B}{SS_W} \times \frac{(N-k)}{(k-1)}$$

Where SS_B is the overall between-cluster variance, SS_W is the overall within-cluster variance, k is the number of clusters, and N is the number of observations.

The overall between-cluster variance SS_B is defined as:

$$SS_B = \sum_{i=1}^{k} n_{i\|m_i - m\|^2}$$

Where k is the number of clusters, m_i is the centroid of cluster i, and m is the overall mean of the sample data.

The overall within-cluster variance SS_W is defined as:

$$SS_W = \sum_{i=1}^{k} \sum_{x \in c_i} \|x - m_i\|$$

Where k is the number of clusters, x is a data point, c_i is the ith cluster, and m_i is the centroid of cluster i.

Typically, well-defined clusters have a large between-cluster variance (SS_B) and a small within-cluster variance (SS_W). As it can be observed in the first equation, a clustering is identified as appropriate by CH, if and only if:

- The within cluster variances are as small as possible
- And, the between-cluster variances are as big as possible
- And, the number of clusters is as small as possible.

Therefore, there is not any cut-off value for CH. This means that the higher the value of CH, the better is the clustering (Maulik, 2002).

2.5.2 Davies–Bouldin

The Davies–Bouldin index, developed by David L. Davies and Donald W. Bouldin in 1979, is a mathematical measure for evaluating clustering algorithms (Davies & Bouldin, 1979). Davies–Bouldin (DB) is an internal evaluation index, therefore the validation of how well the clustering has been done is made by using quantities and features inherent to the dataset (Davies & Bouldin, 1979).

DB is defined based on a measure of good clustering R, similar to the one in CH index, that is a symmetric definition of S and W which are respectively within-cluster and between-cluster similarities. Mathematically speaking $R_{i,j}$ for the i^{th} and j^{th} clusters is defined as (Davies & Bouldin, 1979):

$$R_{i,j} = \frac{S_i + S_j}{W_{i,j}}$$

Accordingly, DB is defined as:

$$DB = {}^{1}\!/_{n} \sum_{i=1}^{n} D_i$$

Where n is the number of cluster and D_i is:

$$D_i = \max R_{i,j} \quad , \qquad i \neq j$$

In other words DB index takes into account both the error caused by representing the data vectors with their cluster centroids (within cluster scatter, S) and the distance between clusters (between cluster separations, W). Thus to make it simple, DB calculates the following ratio:

$$DB \equiv \frac{\text{within cluster scatter}}{\text{between cluster separation}}$$

Therefore a lower DB value indicates a better clustering (Karkkainen & Franti, 2000; Maulik, 2002).

2.5.3 Silhouette

The silhouette value for each point is a measure of how similar that point is to points in its own cluster, when compared to points in other clusters. The silhouette value for the *j*th point, S_j, is defined as:

$$S_j = \frac{b_j - a_j}{max(a_j, b_j)}$$

Where a_j is the average distance from the j^{th} point to the other points in the same cluster as j, and b_j is the minimum average distance from the j^{th} point to points in a different cluster, minimized over clusters.

The silhouette value ranges from -1 to +1. A high silhouette value indicates that j is well-matched to its own cluster, and poorly-matched to neighboring clusters. If most points have a high silhouette value, then the clustering solution is potentially appropriate. On the other hand, if many points have a low or negative silhouette value, then the clustering solution may not be appropriate (Rousseeuw, 1987). In this study an average *S* value of all points has been calculated as a measure of overall clustering quality.

2.6 Association rule mining

Association rule mining (ARM) is a common and well researched method for discovering interesting relations, or so-called associations, between variables in large databases. ARM is capable of identifying interesting association rules by using different statistical measures of interestingness (Piatetsky-Shapiro & Frawley, 1991).

The association rule mining is usually performed on a database which consists of binary transactions with values of either 1 or 0 for each type of feature, which respectively indicate the existence or non-existence of the feature in the given transaction. The ARM looks for the most statistically supported sets of features (itemsets) which most frequently co-exist. Table 1 represents an instant of the binary transaction database used in this study. An association rule is usually represented as an implication of the form $feature\ i \rightarrow feature\ j$ where the $feature\ i$ is regarded as *antecedent* (left-hand-side or LHS) and the $feature\ j$ as *consequent* (right-hand-side or RHS) (Agrawal *et al.*, 1993).

Table 1. An instant of the binary transaction database used in this study. The positive and negative impressions are considered and consequent and the building types as antecedent.

	POS	Neg	Edu	Socialcare
1	1	0	0	0
2	1	0	0	0
3	1	0	1	0
4	1	0	0	0
5	1	0	0	1
6	1	0	0	0
7	1	0	1	1
8	1	0	0	0
9	1	0	1	0
10	1	0	0	1

Several measures are generally used in order to assess and evaluate the rules' strength and quality. However, in this study the two most principal ones, namely *Support* and *Confidence* are used. *Support* represents the occurring frequency of the rule. In other words, it is the proportion of number of transactions containing both X,Y to the whole number of transactions in the database. Moreover, *Confidence* represents the strength of the association by showing, for instance, how often items in Y appear in transactions that contain X (Amelin, 1995).

$$S = \frac{\sigma(X \cup Y)}{Total\ number\ of\ trans.}$$

$$C = \frac{\sigma(X \cup Y)}{\sigma(X)}$$

In a geographical context ARM is considered as a spatial data mining technique and it is sometimes referred to as co-location (Huang, Pei, and Xiong, 2006). Particularly, in a geographical context, ARM is intended to discover possible associations between different georeferenced features which noticeably tend to happen within a certain distance from each other, namely neighborhood distance.

Different algorithms have been proposed and implemented for discovering association rules. However, in this study Apriori algorithm, which is one of the best known ARM algorithms, is used. Apriori uses a 'breadth-first' search strategy to calculate the support of itemsets and it proceeds by identifying the frequent individual items in the database and extending them to larger and larger itemsets as long as those itemsets appear sufficiently often in the database. Finally, the frequent item sets identified by the algorithm are used to determine association rules (Agrawal & Srikant, 1994).

3. Data analysis

The first step which precedes all other tasks in a spatial data analysis project is to explore the data and find its principal characteristics. Knowing the data characteristics helps to determine how to work with it and how to proceed with the analyses. Data analysis is often defined as a "process of inspecting, cleaning, transforming, and modeling data with the goal of highlighting useful information, suggesting conclusions, and supporting decision making" (Adèr & Mellenbergh, 2008). There are many processes associated with data analysis. However, in this research we are mostly interested in studying:

- Characteristics of SoftGIS data,
- distribution of the data,
- general patterns in data,
- and spatial clusters in the data.

In the following sections the above features will be explored and discussed.

3.1 Data distribution

It can easily be observed in Figure 3 that it is highly unlikely for the data to be either evenly or randomly distributed. The distribution of the given data depends on many factors including urban structure, population distribution, public participation, etc.

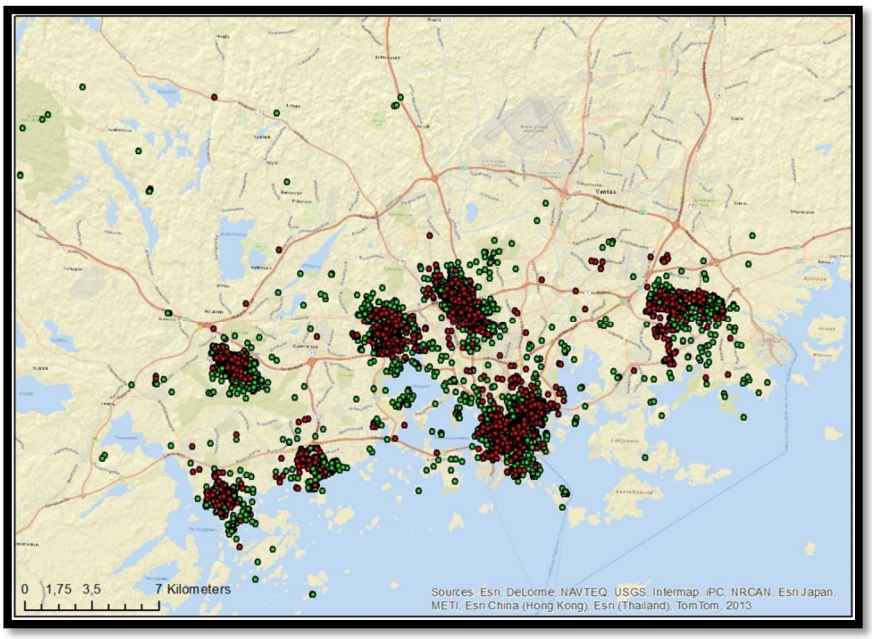

Figure 3. SoftGIS data distributionin in Helsinki metropolitan area. Green and red points represent positive and negative impressions respectively.

We can study the spatial data distribution more precisely by using Moran's Index, also known as Moran's I. Moran's I is a statistical measure of spatial autocorrelation based on both feature locations and feature values simultaneously (Moran, 1950).

In this study spatial statistics toolbox of ArcGIS is used to calculate the Moran's I. The results are presented in Table 2.

Table 2. Moran's I results

Moran's I	**0.335191**
z-score	199.726266

Running Moran's I with the null hypothesis that the data is distributed randomly, results in a large z-score (as in Table 2) which means that is very unlikely for the null hypothesis to be true, therefore the data is highly clustered. Moreover, the positive Moran's I value indicates a positive autocorrelation suggesting that people mostly have similar impressions regarding a certain area.

Although the analysis indicates that generally the points near each other tend to be similar, we can still observe quite many contradicting records within close distances from each other (Figure 3). This is logical since people with different backgrounds and tastes may have different feelings about a certain area. Calculating global Moran's I shows the general distribution of features. Nevertheless, in order to have a deeper insight into the distribution patterns of feature we need to take a closer look at the data. One way of doing so is by calculating Moran's I locally through a process which was first proposed by Anselin (1995).

Anselin Local Moran's I identifies spatial clusters and outliers based on attribute values similar or dissimilar to its surrounding. To do so, the algorithm calculates a local Moran's I value, a z-score, and a p-value. The z-scores and p-values represent the statistical significance of the computed index values (Anselin, 1995). The following criteria are used to interpret the findings:

- A feature is considered to be part of the cluster if its Moran's I is positive and it has a high positive z-score. The P needs to be small enough to indicate the statistical significance of the assumption.
- A feature is considered to be an outlier if its Moran's I is negative and it has a low negative z-score. The P needs to be small enough to indicate the statistical significance of the assumption.

According to these criteria and the calculated values, each feature can be classified and labeled as below:

- PP: A positive feature with positive features in its surrounding (within cluster)
- PN: A positive feature with negative features in its surrounding (outlier)
- NP: A negative feature with positive features in its surrounding (outlier)
- NN: A negative feature with negative features in its surrounding (within cluster)

These classes can be visually illustrated as in Figure 4.

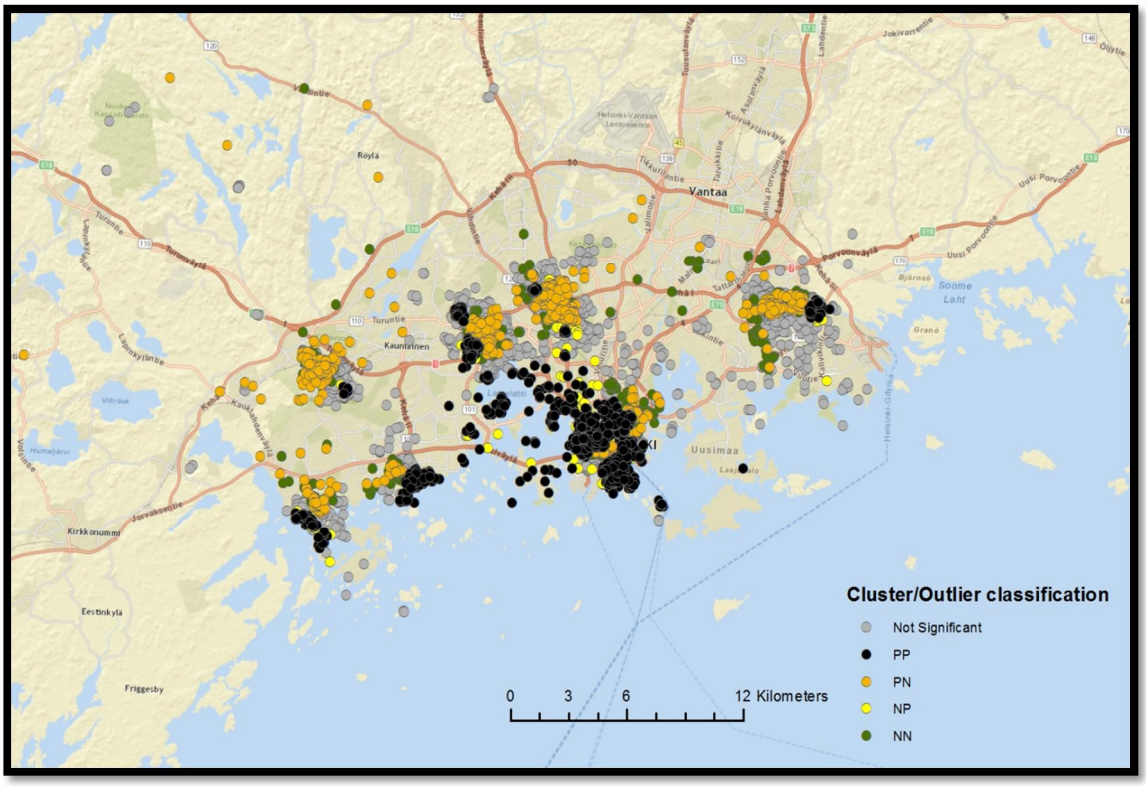

Figure 4. Anselin local Moran's I, Cluster/Outlier classification.

In addition, a directional pattern can be observed within the data as there seems to be more PP features on south western parts (Figure 4). We can check the validity of this observation by exploring the directional effects using correlation coefficients.

Let us represent easting coordinates as X, northing coordinates as Y, and the related impression as *I*. We are interested in calculating the correlation coefficients for the following pairs <X,I> and <Y,I>. The results are presented in Table 3.

Table 3. Correlation coeffcient values

Pair of variables	X & I	Y & I
r	- 0.4085	- 0.2312

The *r* values in Table 3 indicate that as the X and Y values increase, the *I* value decreases. In other words, considering that we have assigned -1 to the negative impressions and +1 to the positive impressions, as we move from south west towards north east, the impressions become more negative (Figure 5).

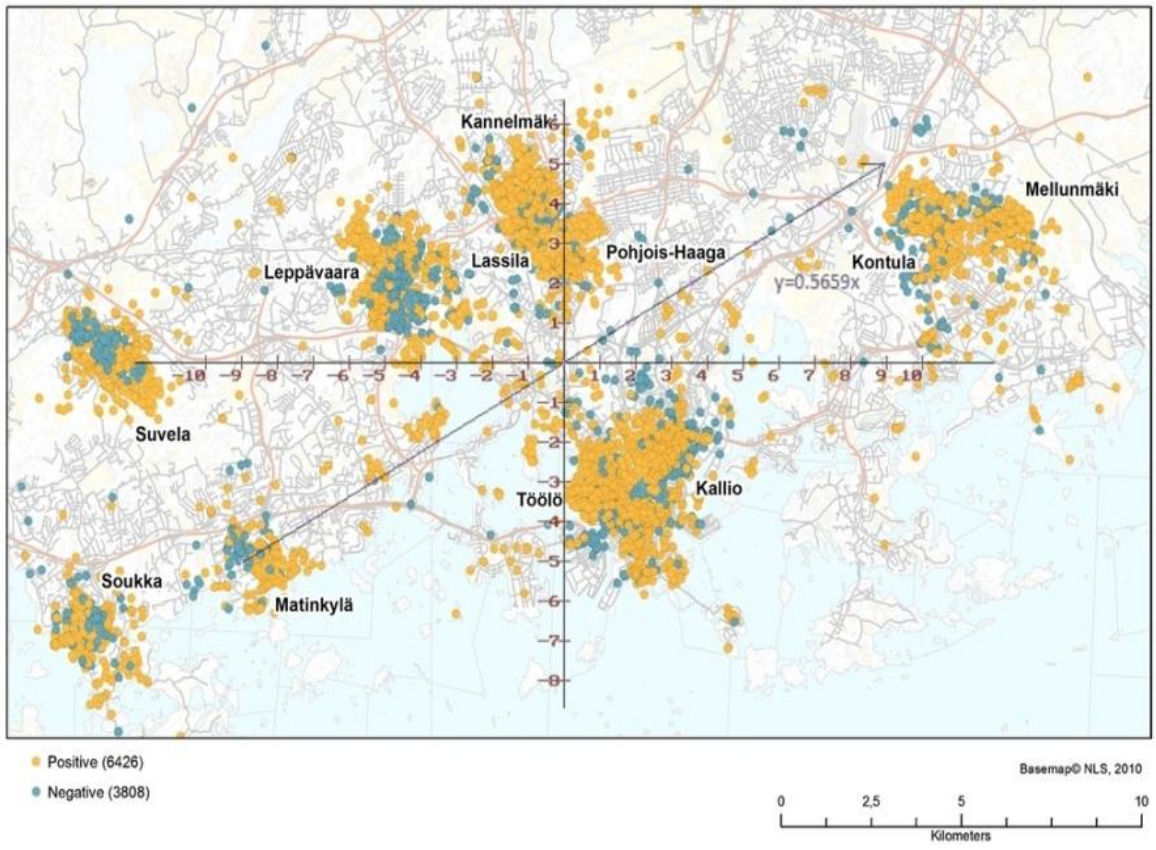

Figure 5. Correlation direction - The arrow represents the direction towards which the overall number of negative impressions rises.

The result is logical as according to the local residents, the most reputable regions of Helsinki region are located on the south west. The eastern areas are typically considered to be of lower reputation.

3.2 Data modeling

The term 'model' has various meanings depending on the area of the study in which it is being used. In spatial sciences model can be defined as a simplified representation of the area of investigation "for purposes of description, explanation, forecasting or planning" (Fotheringham, 2000). In this section we use modeling for the description purpose; therefore the main goal is to identify what kind of model can best describe SoftGIS data.

In actual spatial analysis, no data is perfect. In other words, we always deal with some extent of uncertainty (depending on the quality of the data). Even if one manages to eliminate all systematic errors, the random errors still remain in the data. Random errors cannot be fully eliminated, but they can be measured and mitigated by statistical methods.

SoftGIS data has a distinctive characteristic that distinguishes it from most other kinds of spatial data, in a sense that a point in SoftGIS data does not represent an actual entity and it

may associate with an area of any size and shape. As explained by Kahila (2008), a user clicks on a map and creates a point feature that represents his or her impression of that area. Thus, although the entry is recorded as a point, it does not actually represent a point. Accordingly, in SoftGIS data, not only we deal with the uncertainty which is common to all types of spatial data, we also need to deal with the spatial fuzziness associated with each record. Particularly, the area and shape of influence for each SoftGIS marking has a fuzzy boundary which can affect the analysis reliability.

In other words, in this case there are two main types of uncertainty, namely statistical and cognitive uncertainty. The cognitive uncertainty itself can be further divided into two sub-types: *Vagueness* and *Ambiguity*. *Ambiguity* occurs in situations with two or more alternatives such that the choice between them is left unspecified. *Vagueness* occurs when there is a difficulty in making a precise distinction between objects (Maimon & Rokach, 2010). In SoftGIS data we mostly deal with cognitive uncertainty. One of the best ways of minimizing cognitive uncertainty's effect is by using fuzzy techniques and theories (Maimon & Rokach, 2010). Accordingly, the assumption is that SoftGIS data behaves in a fuzzy way and therefore, fuzzy theory can better describe its behavior. In order to assess the validity of this assumption we try to model the data using two techniques: fuzzy modeling and polynomial modeling. Subsequently we will compare the results and asses the validity of the assumption.

3.2.1 Fuzzy modeling

Fuzzy set theory was introduced by Zadeh (1965) with the aim of overcoming issues caused by classical set theory in dealing with cognitive uncertainties (Maimon & Rokach, 2010). The main difference between classical set theory and fuzzy set theory is in membership definition. In classical set theory, membership is a binary term as a certain element either belongs or does not belong to a set. Fuzzy set theory, on the other hand, permits the gradual assessment of the membership of elements in relation to a set or several sets (Maimon & Rokach, 2010; Zimmermann, 1975).

In this study ANFISEDIT toolbox of Matlab is used to create a fuzzy model of the data. ANFISEDIT toolbox is based on Mamdani fuzzy inference system which maps input characteristics to input membership functions, input membership functions to rules, rules to a set of output characteristics, output characteristics to output membership functions, and ultimately the output membership functions to a single-valued output or a decision associated with the output (Abraham, 2005).

In SoftGIS data the features' distribution follows a pattern that is mainly based on the urban structure and population distribution of the region. Therefore, modeling the whole region would be irrelevant. Accordingly, in order to speculate how the features behave, we study a combination of negative and positive points within a particular cluster rather than the whole region. For this purpose we proceed with Helsinki city center which contains more than twenty five percent of the whole metropolitan area records. The outliers and insignificant values (as described in section 3.1) are excluded from modeling. The input values of the model are the X and Y coordinates and the quantified impression (-1,1) is the output of the model. In order to make the calculations more feasible, the coordinates were normalized (divided by 10^6) so that they would be in the same range as the outputs. Moreover, a random portion of the data was separated in order to be later used for external validation. By trial and error a ten rule optimum fuzzy model was created using Sub-

clustering method (Figure 6). In Mamadani fuzzy inference systems a ten rule model is considered as computationally efficient (Nazemi *et al.*, 2003; Džubur, 2011).

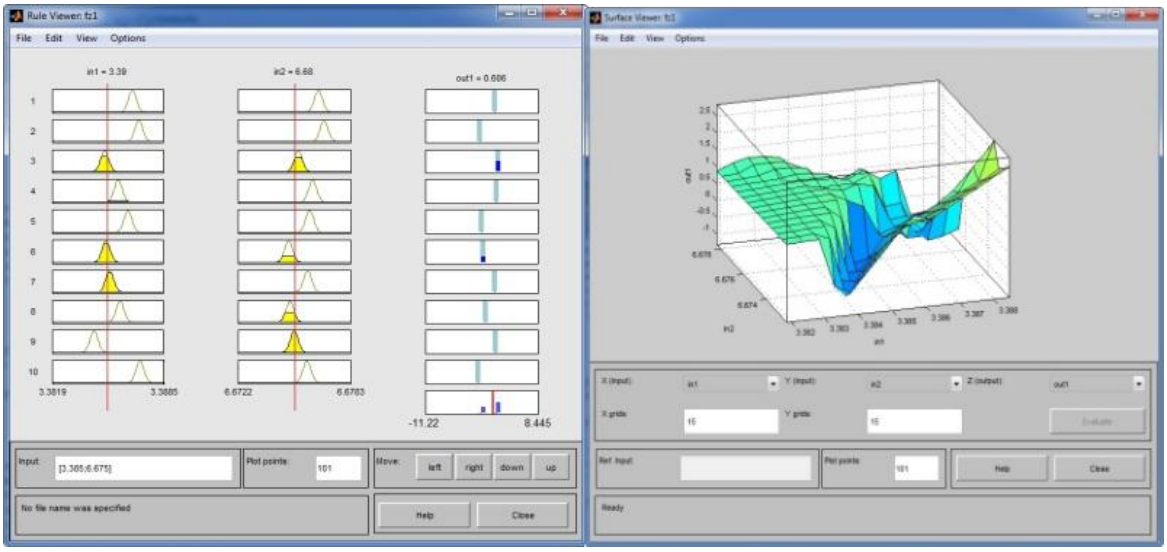

Figure 6. An overview of fuzzy model rule set (left) and the fitted surface (right)

In this study the Matlab's default values were used as the internal parameters, and root mean square error was calculated as a measure of model's quality (RMSE=0.5011).

3.2.2 Polynomial modeling

A polynomial function is one that has the form:

$$y = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0$$

Where n is a non-negative integer that defines the degree of the polynomial. Historically, polynomial models are one of the most frequently used empirical models for curve fitting in statistical modeling (Atkinson, 1989).

The same data as in previous section is used to create a polynomial model to describe the data. Similarly the coordinates were normalized by a division by 10^6 and a random portion of the data was separated in order to be later used for external validation.

It should be noted that in polynomial modeling, the higher the degree of an equation is, the less computationally efficient it would be. For instance a degree one equation is much more convenient to work with than a degree three. That is because high degree equations result in sophisticated computations that are inconsistent to the main purpose of modeling which is to simplify a phenomenon. However, in this case even a double variable polynomial equation with X and Y degrees of five (Figure 7), resulted in a large RMSE (RMSE= 0.7485). This implies that a polynomial function is incapable of describing the irregularities associated with this dataset.

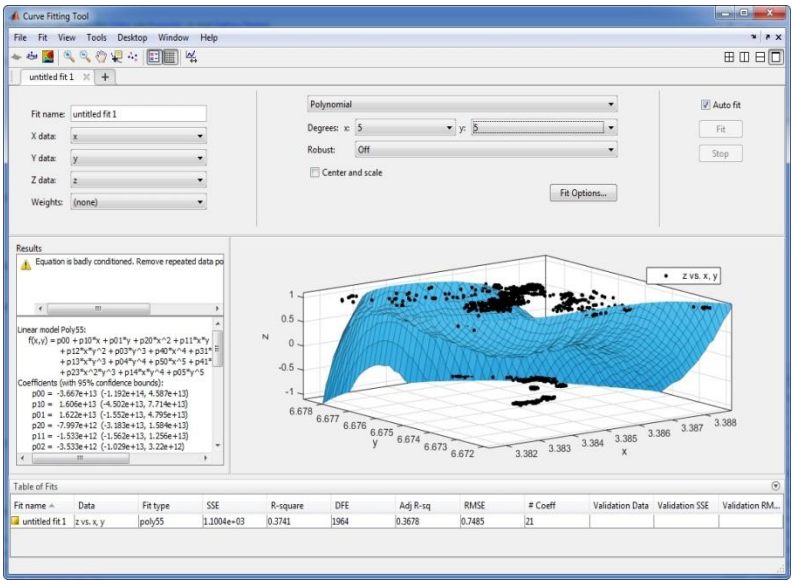

Figure 7. A X^5 Y^5 polynomial model generated by Curve Fitting Toolbox of Matlab.

3.2.3 Conclusion

As stated in section 2.2 the RMSE is a measure of modeling quality validation and is principally formed of the aggregation of differences between model values and the observed values. Therefore, the RMSE values' range corresponds with the output values of the model. In this case the impression value which ranges from -1 to +1 is considered as the output of the model. Accordingly the maximum possible bias of the model from the actual values can be 2 and therefore RMSE values can be any value between 0 and +2. According to this range, the calculated RMSE values for both the fuzzy and the polynomial model are rather high (Table 4). This is a reasonable result based on the high irregularity of the dataset and natural unpredictability of the human related phenomena.

Table 4. RMSE values for fuzzy and polynomial models

Modeling method	Fuzzy modeling	Polynomial modeling
RMSE value	0.5011	0.7485

According to the values presented in Table 4, the fuzzy modeling's RMSE is considerably smaller. With a simple and yet efficient ten rule fuzzy model, the RMSE was calculated as 0.5011 while a complicated degree five polynomial equation yielded a RMSE of as big as 0.7485. Therefore fuzzy modeling has performed more efficiently in describing the dataset.

3.3 Data clustering analysis

Cluster analysis or clustering is the process of grouping a set of physical or abstract objects in such a way that objects in the same group (called a cluster) are more similar to each other than to those in other groups (Miller & Han, 2009). Clustering analysis is one of the main tasks of exploratory data mining and is widely used in numerous applications.

In this study we are interested in finding potential clusters based on their spatial properties. It can easily be observed (as in Figure 3) that the negative and positive points are clustered at several points through the region. Furthermore, the Moran's I analysis in section 3.1 approved the hypothesis that the dataset is highly clustered. In this section mathematical and statistical methods are used in order to investigate the potential clusters. The main purpose of this section is to investigate different clustering methods' potentials in describing the spatial segmentation of the recorded experiences in SoftGIS data.

The methods used in this section can be divided into two major groups: hard clustering and soft (fuzzy) clustering methods. Firstly, three hard clustering methods, namely K-means, SOM and Hierarchical, are applied and compared against each other. There are numerous implementations of these clustering methods; however in this study Matlab implementations are used. The motivation of this section is derived from the hypothesis that according to the fuzzy nature of SoftGIS data and the results obtained in section 3.2 of this report, soft clustering methods can perform better in clustering analysis. Later in this section, in order to check the validity of this hypothesis, the most promising hard clustering method will be compared to a soft clustering method on a mathematical basis which will be discussed.

For the clustering analysis, the positive and negative records are studied separately. Moreover, based on the data distribution's appearance, the potential numbers of clusters for positive and negative points in Helsinki metropolitan area are considered to be eight and seven respectively.

As discussed in section 2.4 there are many clustering validation measures which can help to provide some rough ideas about clustering quality of an algorithm based on the inter-cluster and intra-cluster similarities and variance measures. This study compares clustering methods via three different approaches by using three internal measures: Calinski-Harabasz criterion, Davies–Bouldin index, and Silhouette index.

It should be noted that the clustering results may vary depending on the initial state. Therefore the result values presented in this section have been calculated as the average of ten different algorithm runs.

Table 5 presents the values of Davies–Bouldin (DB) and Calinski-Harabasz (CH) indices for three hard clustering methods, namely K-means, SOM and Hierarchical. For brevity, the averages of values for negative and positive records are presented here.

Table 5. Davies–Bouldin (DB) and Calinski-Harabasz (CH) indices values for three hard clustering methods.

Method	DB	CH
K-means	0.28	15210
SOM	0.4	7943
Hierarchical	0.21	9147

As discussed in section 2.4, CH does not have any cutoff value and it has a direct relation with the quality of clustering; meaning that a higher CH value indicates a better clustering. On the other hand the DB can take any value from -1 to 1 and a lower DB value implies a better clustering. Based on the results presented in Table 5, among the hard clustering methods, K-means yields a considerably higher CH and a rather small DB. This implies that K-means has performed a better clustering and thus, it is chosen as the optimum candidate to be compared with fcm.

Now that we have our best hard clustering candidate, we can compare it with a soft method. Table 6 presents the values of Davies–Bouldin (DB), Calinski-Harabasz (CH), and Silhouette (S) indices for fcm and K-means clustering methods. The positive and negative signs in the table indicate whether the index has been calculated for positive or negative records.

Table 6. Values of Davies–Bouldin (DB), Calinski-Harabasz (CH), and Silhouette (S) indices for fcm and k-means

Method	CH+	CH-	DB+	DB-	S+	S-
K-means	15500	2600	0.49	0.55	0.74	0.87
fcm	16700	15200	0.40	0.57	0.72	0.50

Smaller values of S and DB as well as a higher value of CH indicate a better clustering. Therefore according to Table 6, fcm has a relatively better performance from a mathematical point of view, in comparison to K-means.

3.4 Conclusions

This chapter attempted to study the principal characteristics of the SoftGIS dataset through a set of spatial analysis techniques. The Moran's I results indicate that the distribution of the data is highly clustered within the region.

In order to gain a better understanding of the data distribution, this section attempted to model the dataset using two distinct approaches, namely polynomial and fuzzy modeling. Consequently, the fitted models were compared and the RMSE results indicate that the fuzzy logic has a better capability in describing the SoftGIS data behavior.

Furthermore, the spatial distribution of the dataset was further tested through a clustering analysis. Three hard and one soft (fuzzy) clustering methods were applied to the dataset and the results were assessed by calculating three internal clustering validation measures. The results of these measures imply that the fuzzy clustering method (fcm) has provided a better clustering analysis as it has yielded the highest within-cluster and the lowest between-clusters similarities. Moreover, according to Table 6, the indices imply that the negative records seem to be considerably more fuzzily distributed than positive records. This is in the sense that the measures yielded considerably better values for the negative records in fcm clustering.

In conclusion, the analysis results presented in this chapter approve the hypothesis that the SoftGIS data behaves in a rather fuzzy way and therefore application of the fuzzy logics in studying this data can potentially contribute to better results.

4. Data visualization

SoftGIS data deals with qualitative information. Accordingly, not all its underlying information can be perceived and described through verbal and mathematical descriptions. Moreover, the potential users of SoftGIS data are urban authorities that are not necessarily experts in spatial sciences. These highlight the need for efficient visualization approaches that can provide informative presentation to convey its most important underlying information.

Visualization plays an important role in perception of information. We can often take advantage of our visual perception abilities to amplify our cognition of the abstract data (Shneiderman & Plaisant, 2005). Moreover, through visualization, it is typically easier for the user to understand (Keim, 2001). The most common visualization methods used in geoinformatics are of maps and most people are familiar with them. Furthermore, in geoinformatics many types of thematic maps are also used. A thematic map can be used to emphasize the spatial pattern of one or more geographic attributes (Slocum, 2005). Maps are the primary tool to present, use, interpret, and understand spatial data (Nöllenburg, 2007). Every map communicates a message by emphasizing certain aspects of the underlying data. Moreover, maps highlight interesting information by filtering out unnecessary details of the environment (Nöllenburg, 2007).

Visualization of the SoftGIS data is more challenging than visualizing most other common types of spatial data. That is because the data typically contains huge masses of point, most of which are overlapping. Moreover, existence of many contradicting records in a short distance from each other is common to the SoftGIS data. The reason is that the SoftGIS data represents people's personal experiences and thus it can significantly vary from one person to another.

Therefore, the aim of this chapter is to investigate various visualization methods and assess their capabilities in overcoming these challenges and providing an efficient geographical presentation of SoftGIS data. This chapter first discusses current approaches and their pros and cons and then aims to provide simple, yet reliable, visualizations of SoftGIS data that can facilitate visual analysis. Ultimately, the knowledge discovered through these visualizations will be presented and discussed later in this chapter.

4.1 Assessment of existing methods

Kyttä *et al*. (2013) have presented several methods for visual presentation of SoftGIS data. Each of these methods have capabilities as well has limitations. The simplest method used in the mentioned literature is a point presentation of markings (Figure 8). In this presentation the recorded experiences are divided into two classes of Positive and Negative markings and are demonstrated accordingly.

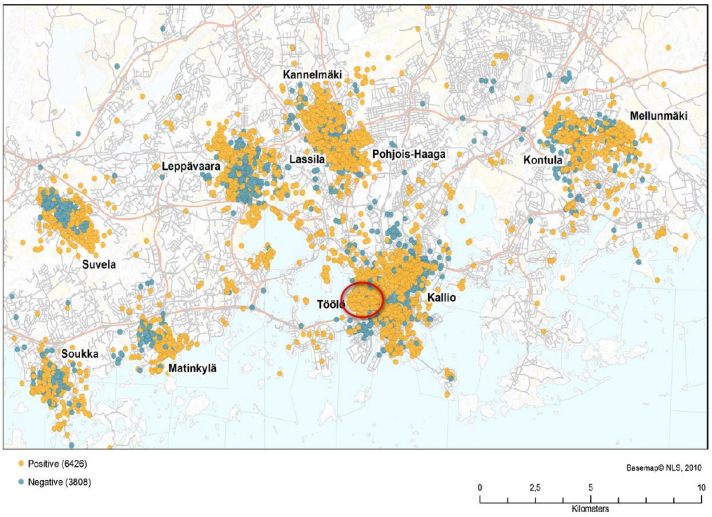

Figure 8. Point presentation of Positive and Negative experiences. From (Kyttä et al., 2013)

As it can be observed in Figure 8, using a point presentation of the recorded experiences has major limitations in conveying the desired information to the user. The first and most significant limitation is that in such a dense dataset it is not very useful to use point representation since there are quite many overlapping features in the areas. For example one might assume that the marked region on the map dominantly contains positive records. However, as illustrated in Figure 9, there are quite many negative records as well, that are not visible in the first presentation.

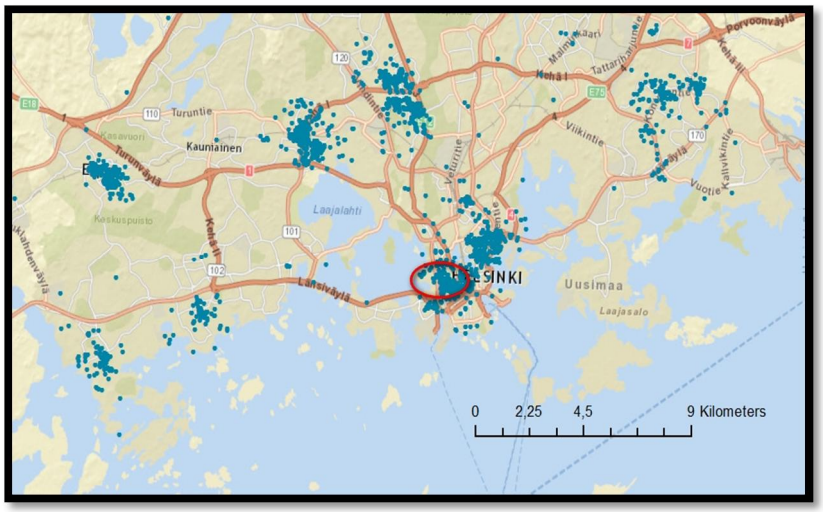

Figure 9. Negative records in Helsinki metropolitan area

Moreover, observing large masses of point clouds does not provide any clear image on the overall impression of the neighborhoods and how it actually varies between different areas.

Another visualization used in the same literature is a ratio map which represents the proportion of positive markings to all markings within a cell of a certain size (Figure 10). Although in this visualization the data is more efficiently managed in comparison to the previous one, it is not yet very practical to use because it requires continual reference to the legend.

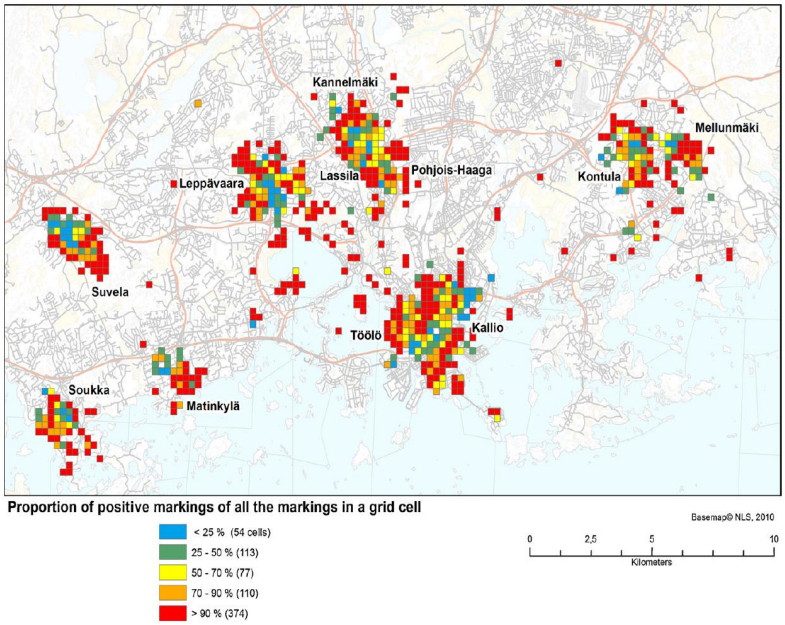

Figure 10. The share of positive and negative place markings (Kyttä *et al*,. 2013).

Another problem regarding this visualization is that it does not either consider or provide any information about the actual amount of negative and positive markings in each cell. This can lead to misinterpretation. For instance, imagine a case where a cell contains only one positive marking and has no negative markings (100% positive). In addition there is one other cell that contains 70 positive and 30 negative markings (70% positive). According to this map, these cells are colored as red and orange respectively. Based on the legend of the map, a red cell in the given map typically implies a more positive impression to the observer than an orange one. However this is not necessarily a true interpretation; that is because according to the records the orange cell has been more frequently reported as positive, than the red cell.

Kyttä *et al*. (2013) propose another visualization method that utilizes a Natural Neighbor interpolation (NN) technique to create a continuous surface of the discrete points (Figure 11).

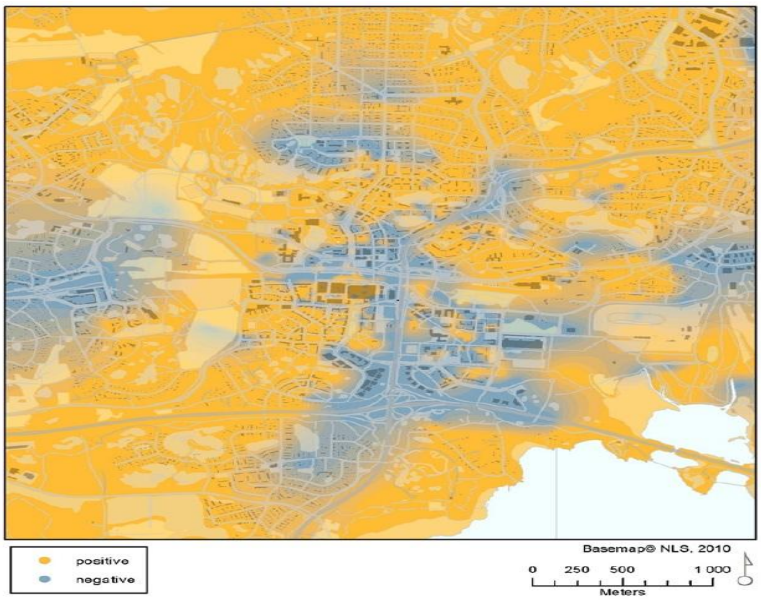

Figure 11. The NN visualization of residents' positive (yellow) and negative (blue) place experiences in Leppävaara area (Kyttä *et al*., 2013).

The interpolation applied in this approach uses the existing observations to make projections and create a continuous surface. This results in a smoother and more visually pleasant presentation of data. However, it has some limitations that may cast doubt on the validity of the interpretations made on this map. In order to explore these limitations, the visualization was recreated in this study (Figure 12).

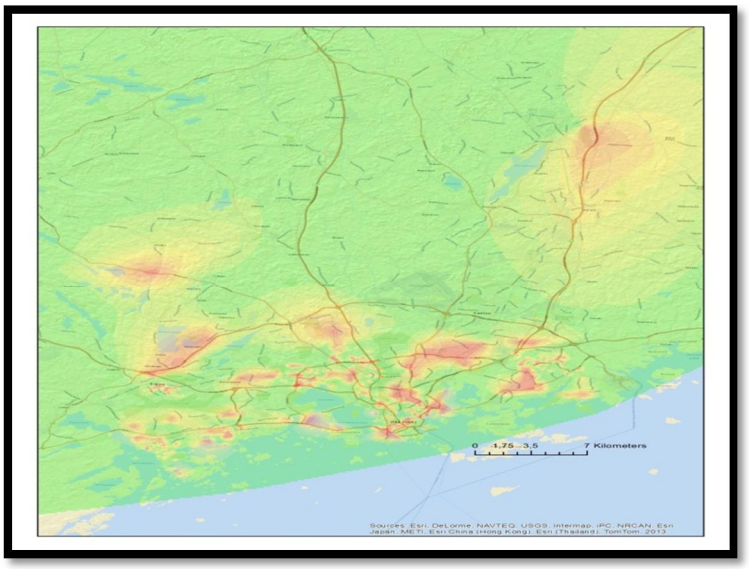

Figure 12. The NN visualization of residents' positive (green) and negative (red) impressions in Helsinki metropolitan area. (Recreated)

The first and most significant limitation regarding this visualization is that it can be misleading. That is because in some cases the prediction may be based on only a few observations (Figure 13). In other words, as a result of interpolation we might have projected values for locations where we do not have enough observations. Moreover, as a result of NN interpolation a contradicting marking (in comparison to the dominant impression in its surrounding) can considerably affect the estimated impression. This casts doubt on how pragmatically this visualization can draw the overall impression of the areas.

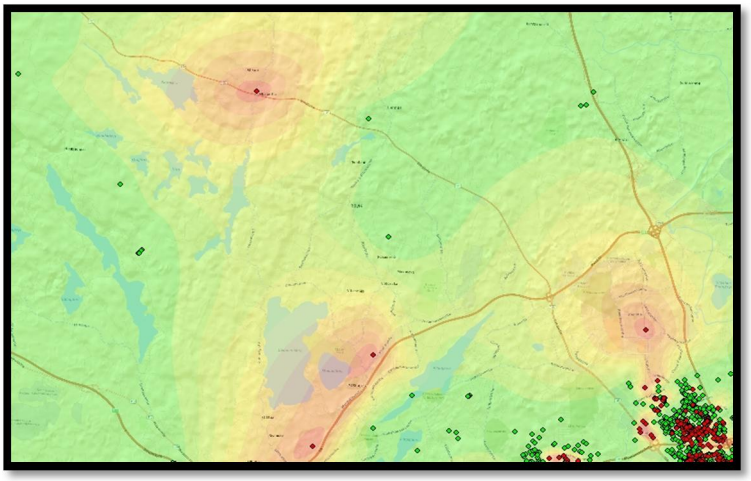

Figure 13. Limitations of using NN interpolation in SoftGIS data visualization. Contradicting markings can cause unrealistic patterns and the predictions may not be well-supported with enough observations.

In summary, it can be observed that the current visual approaches have limitations that can hinder a comprehensive and reliable visual interpretation. Therefore in the coming sections, in order to overcome these limitations, two more visualization approaches are proposed and assessed.

4.2 Point density map

The data consists of large and highly dense clouds of points. Therefore studying the densities can potentially reveal useful information. In order to identify the densest areas in a region, a Kernel density function was applied using ArcMap spatial analyst toolbox. The Kernel density tool calculates the density of features in the neighborhood around those features (Silverman, 1986).

The distances between negative and positive clusters are generally short in this dataset. For example, in the case of Leppävaara area, the negative and positive clusters are located in a close distance from each other, while barely being separated by a railway. Therefore a very long kernel distance (neighborhood distance) would result in oversimplification. Given the fuzzy clustering results from section 3.3, the average distance between nearest pairs of positive and negative clusters in this dataset can be calculated (about 700 meters). Thus a 700 meter kernel distance was considered as an appropriate starting point. Moreover, in

order to have a smoother visualization, a 100 meter cell size was used. The process was done separately on negative and positive datasets resulting in two distinct maps (Figures 14 and 15) depicting negative and positive, so-called, hot spots in the region of study.

In order to evaluate the reliability of these maps and to gain a more realistic understanding of these hot spots we should refer to people's actual perception of these areas. In this case the qualitative descriptions (people's comments), which were provided by the participants, were used to draw the following findings:

- **Leppävaara:** The impression of the residents recorded in this area is polarized. The area to the north of the central railway was experienced very negatively, while the southwest part of the neighborhood was evaluated positively. The negative impressions are mainly associated with this area being the stamping ground for vagrants and drunken people, causing an insecure and fear feeling. On the other hand the positive impressions were mostly associated with the new shopping mall in this area, namely Sello, as well as nice architecture, good transportation accessibility and good facilities.
- **Torkkelinmäki:** This region is generally perceived as positive. The positive impressions were mostly associated with the good social environment and most importantly by the nice park located in that neighborhood.
- **Northern Kamppi:** This region is generally perceived as positive. The positive impressions were mostly associated with the architecture and beautiful scenery.
- **Kinapori:** The area located at north of Torkkelinmäki is generally perceived as negative. The negative impressions were mostly associated with strong feel of insecurity caused by large number of bars, alcohol and drug users, sex shops and Thai parlors.
- **Lehtimäki (Espoon Keskus):** This region is generally perceived as negative. The negative impressions were mostly associated with dirtiness, lack of peace caused by the youth and a nearby school, as well as a feel of insecurity caused by drunks at nights.
- **Southern Malminkartano:** This region is generally perceived as negative. The negative impressions were mostly associated with ugly structures, large number of bars, drunks and addicts.
- **Pohjois-Haaga:** This region is generally perceived as negative. The negative impressions were mostly associated with bars and large number of drunks.

Figure 14. Point density map: Negative hot spots

Figure 15. Point density map: Positive hot spots

4.3 Weighted average visualization

The methods discussed to this point, were incapable of depicting a clear transition between overall impressions within areas. Moreover, for some purposes, such as urban planning, we may require a more precise understanding of the recorded locations. Thus we need to use a more location-based approach that provides a fair representation of the all existing ideas and considers both negative and positive markings simultaneously.

The approach proposed and implemented in this section attempts fulfill the expectations discussed above and tackle the aforesaid limitations by calculating a weighted average of markings within a predefined cell. The idea of using an average, is in response to the need to draw a fair overall impression of various experiences in an area, and is driven by the ability of *averaging* to provide a fair representative of various quantities.

Mathematically speaking, we are interested in calculating I (Impression) for every given cell as below:

$$I = \frac{n_n \times negative + n_p \times positive}{n_n + n_p}$$

Where n_n and n_p are the number of respectively negative and positive markings within the cell and negative and positive values are defined as the -1 and +1 respectively. Obviously, I is a continuous value in the range of [-1, 1] and the higher the value, the more positive the overall impression would be.

The size and shape of the cells can considerably affect the visualization result. Obviously the larger the size of the cell, the less locationally precise it would be, though the visualization would become smoother. It should be noted that while choosing the cell size, the desired scale of presentation should also be taken into account. In this case and according to the scale, a 300 meter cell size revealed the best level of details in comparison to other tried sizes. Ultimately the calculated I's were visualized on a gradual scale as in Figure 16.

A closer look to this map indicates that the presented patterns conform to the textual comments received from the participants. For instance in case of Leppävaara area, it was frequently reported that the eastern side of the highway is perceived negative while the western side is often regarded as positive. This trend has been more efficiently captured in this visualization (Figure 17) in comparison to previous approaches (Figures 10, 11, 14, and 15).

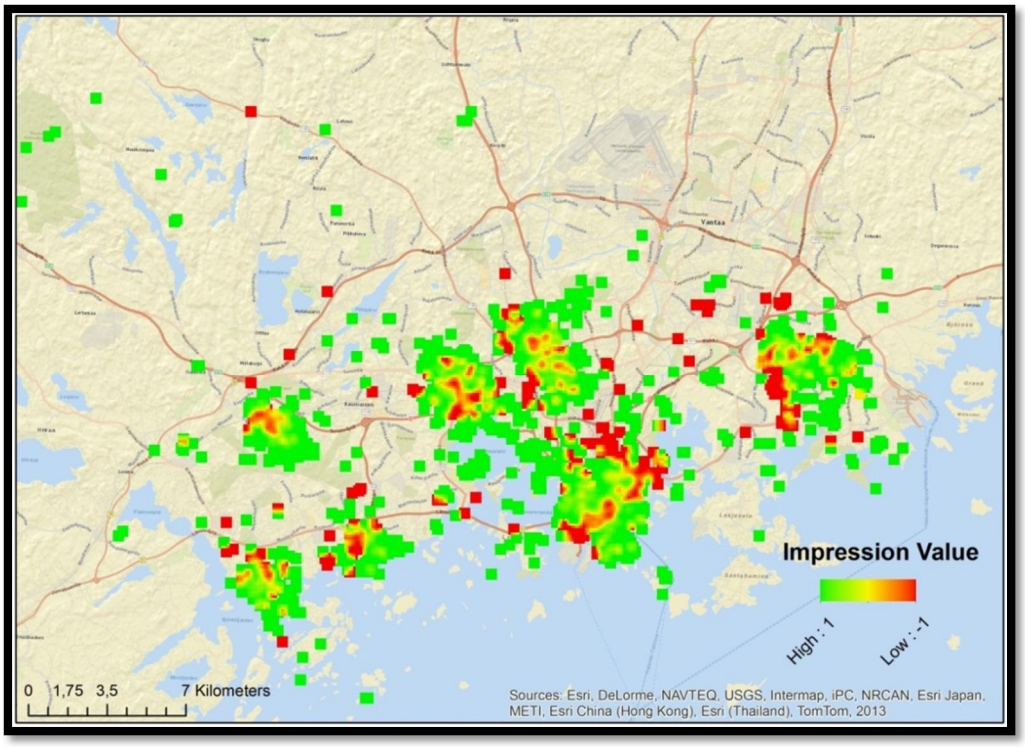

Figure 16. Weighted average impression map (WAV)

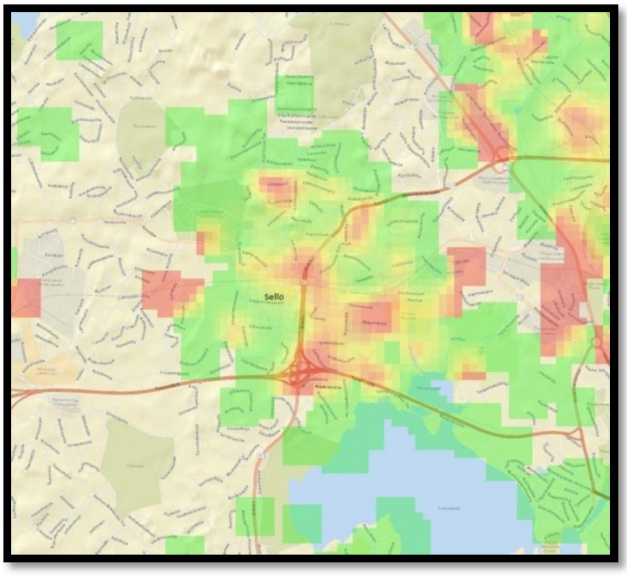

Figure 17. Negative and positive impressions in Leppävaara area (green and red colors represent positive and negative impressions respectively)

4.4 Conclusions

Using an appropriate visualization can both improve and facilitate the visual interpretation of the spatial data. SoftGIS data has certain characteristics that make its visualization more demanding than other common types of spatial data. Most outstandingly, the existence of large crowds of points and contradicting features are the most important challenges in visualizing this data. That is because people have different tastes and therefore different perceptions of the environment. Thus in each crowd of points different ideas can be observed (positive and negative). For instance in a region such as Lehtimäki, adults complain about the noisy and rebel youth causing a negative impression; whereas, obviously the youth find the very same region as friendly and cozy.

This chapter explored the potentials of different visualization approaches in facilitating visual interpretation of SoftGIS data. The existing methods have limitations that result in less informative maps or even confusion of the user. These methods are very sensitive to contradicting markings, they do not indicate the density and amount of observations, and they do not provide a clear image of the overall impression and its transitions throughout the region of study. Therefore in order to tackle these limitations two other visualizations were applied in this study, namely point density map and Weighted Average Visualization (WAV).

A point density map can identify the negative and positive hot spots in the region and provide a visual understanding of the amount and density of observations in different areas. This visualization provides useful information by both considering and illustrating the amount and density of observation throughout the region. However, this visualization approach is mostly suitable for pointing out the gist of patterns and it is incapable of spotting more detailed trends. In addition, depiction and analysis of negative and positive experiences in two separate maps typically lessens the efficiency of the visualization in conveying the information to the observer. This also hinders depiction of a realistic overall impression of areas.

Furthermore, the weighted average visualization is capable of revealing patterns which were not visible in other visualizations. For instance, in Helsinki city center, south of Kamppi (Figure 16), a highly negative impression can be observed. This is mostly associated with city center traffic congestions and dirtiness according to participants' views. In addition, this visualization preserves the locational quality. Therefore, for each location the overall satisfaction of the residents can be extracted. This overall satisfaction is based on the average of negative and positive impressions within the area. Moreover, WAV is capable of capturing the gradual transition of impressions between different areas. This contributes to a smoother and more realistic image of the region. Nevertheless, there is this limitation that it does not provide any information about the actual amount of observations for each cell.

In conclusion, each visualization approach has some advantages and disadvantages and there is no absolutely comprehensive way of presenting the data. Suitability of visualization methods is a relative concept and should be assessed in accordance to the desired outcome and the type of knowledge we are looking for. Table 7 presents an overview of the aforesaid capabilities discussed in this chapter.

Table 7. A brief comparison of different visualization methods

Visualization Method Quality	Point map	Ratio map	NN	Density map	WAV
Attractiveness		X	X		X
Preserving locational quality	X	X			X
Capturing transitions					X
Effective data management		X	X		X
Robustness towards outliers and anomalies	X	X		X	X
Informativeness and reliability of results		X		X	X
Depiction of amounts of observations	X			X	
Providing a fair image of all opinions			X		X
User friendliness			X		X

5. Association rule mining

Similar to many research fields, with the recent advancements in technology and the enrichment of spatial data acquisition in different public and private sectors, geography and spatial sciences have become more computation-rich (Miller & Han, 2009). Using these computational techniques can help spatial scientists to make more profound and concrete discoveries while studying geocoded and spatial datasets.

In previous chapters interesting discoveries were made through using different visual and analytical exploratory techniques. The purpose of this chapter is to reach a new, and perhaps more profound, level of knowledge discovery within Helsinki region urban impression dataset, by using a statistically supported spatial data mining technique.

The studies done in previous chapters imply a considerable association between people's perception of their urban environment and the buildings and properties in their surroundings. Therefore this chapter attempts to mine possible associations between the recorded impressions and the building types using Apriori algorithm.

5.1 Implementation

As described in section 2.6, association rules provide a very simple but useful form of rule patterns for data mining (Hand, 2001). One of the important factors in spatial association rule mining (co-location) is the definition of neighborhoods. It is based on these neighborhoods that the Apriori algorithm can determine whether two features are associated or not. Therefore the shape and size of the objects' neighborhoods can considerably affect the outcomes of the algorithm.

According to Karasova (2005) using grids for the neighborhood definitions has certain limitations that make them unsuitable for this study. For example two very close features may not be identified as associated just if they happen to be located at different grid cells (Figure 18).

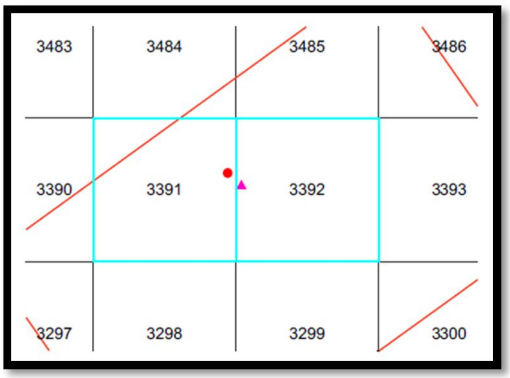

Figure 18. The grid neighborhood problem on the edges of a cell (Karasova, 2005).

Accordingly in this case, a circle buffer is a more suitable neighborhood definition. Working on the same building data as in (Karasova, 2005), a 50 meter radius was identified as the optimum buffer distance. However, in this case there is an extra source of uncertainty which is caused by the SoftGIS data's cognitive uncertainty and should be taken into account. Given the fuzzy clustering results from section 3.3, the standard deviation within each cluster can be calculated. The average standard deviation for all clusters approximately equals to 35 meters; therefore a 35 meter addition to the buffer distance can be a good starting point for considering the existing cognitive uncertainty in the given dataset. Ultimately, a 85 meter circular buffer was applied as the neighborhood around each positive and negative marking (Figure 19) and a unique ID number was assigned to each of them.

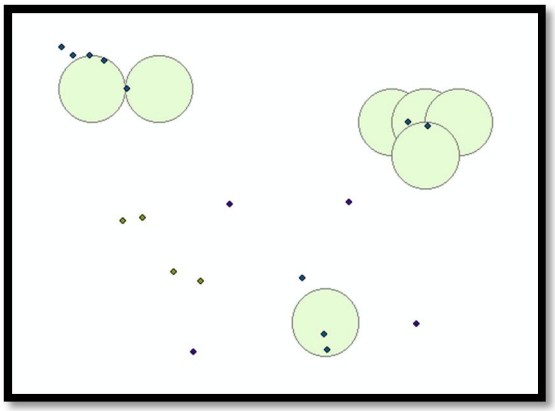

Figure 19. 85 meter buffers around markings. The points represent the buildings.

The building type dataset consisted of many classes. In order to narrow than results to more specific and potentially interesting ones, the dataset was reclassified to the following twelve classes using the discoveries made in chapter 4:

Educational: containing universities, schools and libraries

Social Care: nursing homes, prisons, penitentiaries, day care centers

Holiday: cottages, hotels and accommodation facilities

Residential: residential buildings

Utility: maintenance facilities, heating stations, power plants, parking lots, car maintenance services

Sauna: Sauna buildings

Office: office buildings

Cultural: theaters, sport facilities, religious buildings

Agricultural and industrial: containing agricultural and industrial buildings

Transport: containing transportation building such as metro stations, train stations etc.

Shops: containing shopping centers and malls

Restaurant and bar: containing restaurants and bars

It should be noted that Saunas are of significant cultural influence in Finland and therefore they have been specified as a separate class in this study.

After pre-processing the dataset, a spatial join tool was used to identify the intersecting features. It should be noted that the output of this process is not directly usable in Apriori algorithm. Thus, the table needs to be processed and converted into binary transactional format, which was done in this study by making some python coding (Appendix B).

Finally the Apriori algorithm was applied using IBM SPSS Modeler and the interesting association rules were extracted as represented in Table 8. Since the dataset is large and there are twelve classes, it makes sense that in most cases the support value is small. Accordingly, the confidence has majorly been considered as the measure of the association rule strength and validity.

Table 8. Discovered association rules

Impression (Consequent)	Building type(Antecedent)	Support %	Confidence %
Positive	Residential	20.437	90.762
Positive	Sauna/residential	3.014	97.802
Positive	Sauna	3.942	96.639
Positive	Restaurants & bars	1.722	90.385
Positive	Shops	8.91	59.48
Negative	Transport	2.352	84.444
Negative	Office	43.524	52.74
Negative	Social care	5.035	67.105
Negative	Restaurant & bars/residential	0.331	70
Negative	Utility	19.212	59.31
Negative	Agricultural & industrial	3.279	66.667
Negative	Agricultural &	0.132	100

	industrial/social care		

5.2 Interpretation of results

Although the association rules illustrated in Table 8 are rather self-explanatory, they can be even more informative when interpreted considering the cultural/urban context of the region of the study. Accordingly following interpretations can be made:

- People typically tend to like their residential area. Accordingly more than 90 percent of markings made in residential areas represent a positive impression.
- Saunas are very important and adored in Finnish culture. Thus, residential areas with a sauna building nearby (homes with sauna) are even more desirable to Finns.
- Although many negative comments were received regarding bars and restaurants in Helsinki region, the results imply that they are typically perceived positively in more than 90 percent of cases. However, when the bars and restaurants are located within residential areas they generate a negative impression (with a confidence of 70 percent). This makes sense as they typically disturb the peace in residential areas.
- The majority of shopping centers generate a positive impression for the habitants (about 60 percent). However, there are exceptions. For instance, in northern Leppävaara region (as discussed in section 4.2) there is an old shopping center built in mid-80s, which nowadays hosts some pubs, flea markets and gym clubs (Kyttä *et al.*, 2013).
- Social care buildings mostly generate a negative impression in their surrounding (with a confidence of almost 67 percent). This is a logical statement as prisons, orphanages and penitentiaries etc. are typically perceived negatively.
- Utility, agricultural and industrial buildings generate a negative impression in most cases. This is a logical statement as these types of properties typically cause disturbances to their surroundings.
- Agricultural, industrial and social care properties rarely co-exist in an area (support 0.132 percent). However, when they do, they interestingly generate a negative impression with a 100 percent confidence.

5.3 Conclusions

Using spatial data mining techniques helps to discover the less explicit information in large datasets. This study, by using the association rule mining, managed to find associations and patterns which were not explicitly found by other visual and analytical techniques. Another advantage of using ARM is that any discovery through this technique is supported by statistical measures, namely support and confidence values, which can to some extent, provide an overall insight of the validity of the findings. However it should be noted that in large datasets, same as the one used in this study, the value of support for the discovered rules is typically small. That is because, for instance, although there are many office buildings in Helsinki region, they are still a small fraction of the whole number of

buildings in this area. Accordingly any rule containing offices will normally yield a rather small support value.

Moreover, the neighborhood distance should be carefully chosen as it can significantly influence the discovered rules. In this study, the average standard deviation of clusters was taken into account as a measure of locational uncertainty. This consideration contributed to better results in comparison to other assumptions tried.

6. Conclusions

SoftGIS is an innovative attempt in acquisition of geocoded (hard) qualitative (soft) data that is designed to generate spatio-qualitative datasets through map-based surveys. These datasets are usually large and typically demand efficient spatial techniques to reveal useful information. Typical spatial analysis provides techniques for discovering patterns from large geographical datasets. However, due to distinct characteristics of this type of data, these techniques should be modified and used strategically in order to reach concrete knowledge.

This study concerns the process of knowledge discovery within SoftGIS dataset with the aim of providing an appropriate methodology in spatio-qualitative data analysis and supporting the role of such datasets in providing subjective information. The study presents different approaches while paying special attention to the characteristics of the data being studied. While performing a case study of urban impression in Helsinki, this thesis applied various methods which are originated from different fields of spatial sciences and it attempts to provide the reader with an inclusive viewpoint, one that encompasses different aspects. Accordingly, the study documents all of these aspects in order to provide guidelines to the process of spatio-qualitative data knowledge discovery for its future users.

6.1 Methods

The study brings a broad viewpoint on the process of knowledge discovery within spatio-qualitative datasets by applying consecutive methods that vary from visualizations, through statistical approaches, to computational approaches. More specifically, the study uses a sequence of data mining, exploratory and statistical analysis techniques which are organized as a four stage knowledge discovery process in order to reach the desired knowledge (Figure 20).

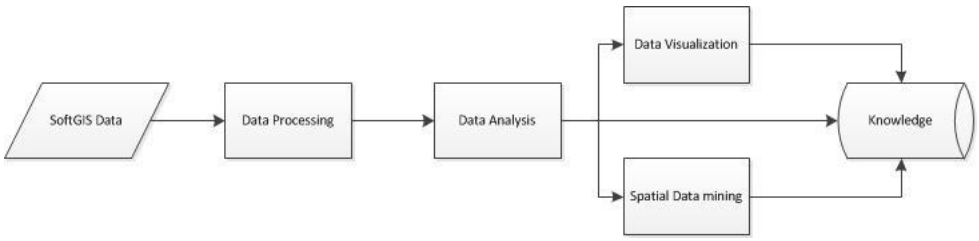

Figure 20. Brief presentation of the four stage knowledge discovery process used in this study

First, the process begins with a *Data processing* stage which is intended to prepare the dataset by data cleaning, classification and uncertainty analysis. Moreover, in order to facilitate the processes and enable the use of mathematics, the SoftGIS data was quantified. In addition, in this stage the potential sources of uncertainty were explored. In this study the data processing stage has been implicitly presented within introduction and data analysis parts.

Second, *Data analysis* stage attempts to identify the principal patterns and characteristics of the dataset through a range of statistical and computational methods. Particularly, through this stage, the SoftGIS dataset's characteristics were explored and the findings were later taken into account in the other two stages. In this stage, Moran's I and several clustering methods were applied to gain a better understanding of the spatial distribution of the data and a correlation coefficient was used to discover the existing directional effects. Moreover, the potentials of using fuzzy logic were thoroughly investigated in both modeling and clustering analysis of the data.

Third, the *Data visualization* stage aims to provide the users with useful information via visual conception. However, since the study involved qualitative data and due to the specific characteristics of SoftGIS data identified in the previous steps, this stage was more challenging than usual. The current visual approaches discussed in the existing literature have certain limitations that signify the need for a more reliable and informative approach which is the main purpose of this step. Accordingly, appropriate usage of Kernel density map and implementation of a novel visualization technique, namely weighted average visualization, resulted in revelation of interesting information.

Finally, by considering the discoveries made in the previous stages, use of appropriate spatial data mining techniques, with an aim of making more profound discoveries, was facilitated. Accordingly an association rule mining process was performed that revealed interesting associations between the recorded impressions and the surrounding building types.

6.2 Summary of results

A closer look at the SoftGIS data and its acquisition procedure highlights the considerable existence of an extra source of uncertainty, namely cognitive uncertainty. Cognitive uncertainty is a side effect of human recognition abilities and results in ambiguity and vagueness of shape, size, and exact location of each feature. Therefore in this study different criteria were used in order to identify the most promising approach in ameliorating this uncertainty. The data modeling results implied that the fuzzy logic has by far a better capability of describing SoftGIS data in comparison to conventional polynomial models. This claim was supported by calculation of root mean square errors for each type of model. Similar findings were made in clustering analysis as the fuzzy c-means clustering method yielded better values of internal cluster validation measures in comparison to other conventional methods such as K-means.

In addition, various visualization methods were discussed in this study. Each of these visualization techniques has certain capabilities as well as some limitations. On whole, the results imply that the weighted average visualization (WAV) proposed in this study has better capabilities in comparison to the other methods. Yet it should be noted that there is no such a thing as an absolutely comprehensive visualization method. Propriety of visualization approaches is a relative concept and therefore should be assessed in accordance to the data, context, and the type of knowledge one is interested in.

Ultimately, considering the context of study can provide us with clues on potential reasons behind certain behavioral observations of the phenomenon. For instance, in case of urban impression, given the fact that the context of study is an urban environment, buildings can

potentially play an important role in people's impressions. Moreover, the visualizations imply a considerable association between the types of buildings in an area and the recorded impressions. The association rule mining applied in this study approved this claim and discovered interesting associations. A summary of these association rules can be observed in Figure 21.

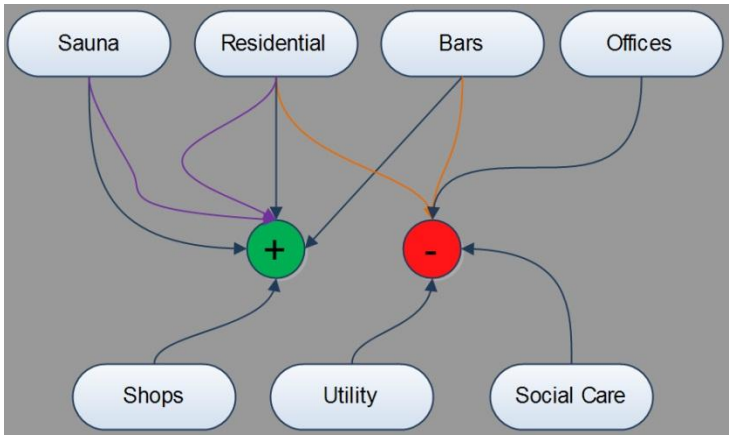

Figure 21. A summary of the discovered associations between building types and impressions. Black arrows represent singular antecedents and the other same-color arrows indicate pair antecedents.

As it can be speculated in Figure 21, most positive impressions are associated with residential buildings, shops, saunas, and bars. However, when bars and residential buildings co-locate, they generate a negative impression. Moreover, the impressions associated with offices, utility, and social care buildings are generally negative.

6.3 Significance of the results

Most of the studies done on SoftGIS address its acquisition procedure and potential applications. This leaves the SoftGIS data analysis and knowledge discovery a rather untouched problem. Moreover, there have not been any studies on the characteristics of this data and how it should be treated. However, there have been several attempts in visualization of SoftGIS data that merely consider the characteristics of the dataset and are incapable of providing a thorough understanding of the data (Kyttä *et al.,* 2013; Kahila, 2008). Kyttä *et al.* (2013) propose a ratio map of the SoftGIS data that neglects the amounts of observation in different cells and can cause misinterpretation. In the same literature, natural neighbor interpolation technique was implemented to provide a smooth image of urban experiences in the region. This visualization oversimplifies the locational quality of the dataset and is vulnerable to the contradicting records common to the SoftGIS data. This can result in misrepresentation of the information in some cases. Similarly, use of inverse-distance weighting interpolation (IDW) in (Kahila, 2008) has analogous limitations and can result in misrepresentation of the information.

In addition, the existing literatures have overlooked the computational capacity of the SoftGIS data. This has led to the negligence of applications of computational and data mining methods in SoftGIS data analysis.

This thesis analyzed the SoftGIS data and contributed to a better understanding of its principal characteristics. Moreover, it explored the potentials and limitations of different visualization methods and highlighted the type of information each of them can reveal. In addition, this study signified the possibility of association between the SoftGIS data and other layers of information and how they can contribute to knowledge discovery when analyzed in a common context. In a big picture, this thesis provided sample criteria of how the SoftGIS data should be treated and it presented a paradigm of SoftGIS knowledge discovery that can be considered as a cornerstone of future studies.

Consequently, the methods and hypotheses introduced in this thesis are not necessarily limited to SoftGIS datasets. The presented ideas can be extended and generalized so that they would be applicable to any spatio-qualitative dataset of similar nature and characteristics.

6.4 Limitations and further research

SoftGIS data analysis is a rather new topic and requires further research. In this study, a fuzzy characteristic of SoftGIS datasets was identified and potentials of fuzzy methods were explored and highlighted in working with this dataset. However, this hypothesis requires to be tested on different datasets and in different regions in order to be considered as a profound theorem. Moreover, in this study the simplest implementations of fuzzy clustering and fuzzy modeling were used. The results can potentially be further improved by applying more advanced implementations of these techniques. Furthermore, in several parts of this study the cell shapes and buffers were considered as squares and circles respectively. A more rigorous study of the case context can help us to identify better definitions of the neighborhoods. This can potentially result in better and more precise findings.

As stated several times in this study, the fuzzy methods proved to provide better solutions to the modeling of SoftGIS data distribution and cluster analysis. This can open a new scope for further research as this idea can be expanded by e.g. creating a fuzzy visualization of the dataset or using a fuzzy-apriori algorithm for ARM (Buczak & Gifford, 2010; Hüllermeier, 2005). Taking these approaches can potentially improve the mitigation of the cognitive uncertainty's effects and enhance the reliability of the discoveries.

Consequently, the findings of this study accompanied with other researches can lead to the creation of an inclusive spatio-qualitative analysis toolbox which can ease and speed the analysis of this type of data. Such a toolbox should potentially consist of various data analysis, visualization, and data mining tools and methods that can facilitate the process of knowledge discovery within spatio-qualitative data.

In summary, the results of this study motivate further research for the rigorous analysis of SoftGIS data and spatio-qualitative data in general. This study highlights the computational capacity of these datasets and its applications in exploratory and visual presentation of the

data. This research presents a paradigm of spatio-qualitative data knowledge discovery and thus has a practical impact on future studies involving such datasets.

References

Abraham, A. 2005. Adaptation of fuzzy inference system using neural learning. In *Fuzzy Systems Engineering.* Springer Berlin Heidelberg. pp. 53-83.

Ade`r, H., Mellenbergh, G. and Hand, D. 2011. *Advising on research methods: a consultant's companion*. Huizen: Johannes van Kessel Publishing.

Agrawal, R., & Srikant, R. 1994. Fast algorithms for mining association rules. In Proceedings of. *20th int. conf. very large data bases, VLDB*. September, Vol. 1215, pp. 487-499.

Agrawal, R., Imieliński, T., & Swami, A. 1993. Mining association rules between sets of items in large databases. In *ACM SIGMOD Record,* 22(2), pp. 207-216.

Ahmed, M. N., Yamany, S. M., Mohamed, N., Farag, A. A., & Moriarty, T. 2002. A modified fuzzy c-means algorithm for bias field estimation and segmentation of MRI data. Medical Imaging, IEEE Transactions on, 21(3), pp. 193-199.

Amelin, L. 1995. Local indicators of spatial association: LISA. *Geographical Analysis*, 27(2), pp. 93-115.

Anselin, L. 1995. Local indicators of spatial association—LISA. *Geographical analysis*, *27*(2), 93-115.

Atkinson, K. E. 1989. *An introduction to numerical analysis*. New York: Wiley.

Buczak, A. L. & Gifford, C. M. 2010. Fuzzy association rule mining for community crime pattern discovery. p. 2.

Costa, J. A. F. 2010. Clustering and visualizing SOM results. *Springer*, pp. 334-343.

Davies, D. L. & Bouldin, D. W. 1979. A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (2), pp. 224-227.

Dowdy, S. M. & Wearden, S. 1983. *Statistics for research*. New York: Wiley.

Dunn, J. C. 1974. Well-separated clusters and optimal fuzzy partitions. *Journal of cybernetics*, 4(1), pp. 95-104.

Džubur, Ž. 2011. Assessment of factors of ecological acceptability as a criterion when deciding on construction material, In Proceedings of *Trends in the Development of Machinery and Associated Technology*, Prague, Czech Republic, 12-18 September. Prague: pp. 445-448.

Färber, I., Günnemann, S., Kriegel, H. P., Kröger, P., Müller, E., Schubert, E., ... & Zimek, A. 2010. On using class-labels in evaluation of clusterings. In *MultiClust: 1st International Workshop on Discovering, Summarizing and Using Multiple Clusterings Held in Conjunction with KDD* July 25-28.

Fotheringham, A. S. and Wegener, M. 2000. *Spatial models and GIS*. London: Taylor & Francis.

Guttman, L. 1944. A basis for scaling qualitative data. *American sociological review*, pp. 139-150.

Hand, D. J., Mannila, H. and Smyth, P. 2001. *Principles of data mining*. Cambridge, Mass.: MIT Press.

Hsy.fi. 2014. *HSY-Geographical information*. [online] Available at: http://www.hsy.fi/en/regionalinfo/urban/gis/Pages/default.aspx [Accessed: 14 Feb 2014].

Huang, Y., Pei, J. and Xiong, H. 2006. Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, 10(3), pp. 239-260.

Hüllermeier, E. 2005. Fuzzy methods in machine learning and data mining: Status and prospects. *Fuzzy Sets and Systems*, 156(3), pp. 387-406.

Hyndman, R. J. and Koehler, A. B. 2006. Another look at measures of forecast accuracy. *International Journal of Forecasting*, 22(4), pp. 679-688.

Johnson, S. C. 1967. Hierarchical clustering schemes. *Psychometrika*, 32(3), pp. 241-254.

Kahila, M. 2008. Possibilities of Web-based softGIS Method in Revealing Residents Evaluation Knowledge of the Living Environment. In *FUTURE-Future Urban Research in Europe, The Electronic City Conference, Bratislava*.

Kahila, M. and Kyttä, M. 2011. Web-based SoftGIS method in research and urban planning practices. *The Electronic City*, 1 p. 199.

Kahila, M., & Kyttä, M. 2006. The Use of Web-based SoftGIS-method in the Urban Planning Practices. In *Proceedings of the Conference on Urban Conditions and Life Changes*.

Kahila, M., & Kyttä, M. 2009. SoftGIS as a bridge-builder in collaborative urban planning. In *Planning support systems best practice and new methods, Springer Netherlands*, pp. 389-411.

Karasova, V. 2005. *Spatial data mining as a tool for improving geographical models*. Master's thesis. Helsinki University of Technology.

Karkkainen, I. and Franti, P. 2000. Minimization of the value of Davies-Bouldin index.

Kaymak, U., & Setnes, M. 2000. *Extended fuzzy clustering algorithms* (No. ERS-2000-51-LIS). Erasmus Research Institute of Management (ERIM).

Keim, D. A. 2001. Visual exploration of large data sets. *Communications of the ACM*, 44(8), pp. 38-44.

Klawonn, F. 2008. Fuzzy clustering: Insights and a new approach. *Mathware & soft computing*, *11*(3), 125-142.

Koperski K. and Han J. 1995. Discovery of Spatial Association Rules in Geographic Information Databases, Proceedings of *4th International Symposium on Large Spatial Databases*, pp. 47-66.

Koperski K., Adhikary J. and Han J. 1996. Spatial Data Mining: Progress and Challenges, in IGMOD96 *Workshop on Research Issues on Data mining and Knowledge Discovery*.

Kyttä, M., Broberg, A., Tzoulas, T., & Snabb, K. 2013. Towards contextually sensitive urban densification: Location-based SoftGIS knowledge revealing perceived residential environmental quality. *Landscape and Urban Planning*, *113*, pp. 30-46.

Leung, Y. 2010. *Knowledge discovery in spatial data*. Heidelberg: Springer.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* 1(281-297), p. 14.

Maimon, O. Z. and Rokach, L. 2010. *Data mining and knowledge discovery handbook*. New York: Springer.

Manning, C. D., Raghavan, P. and Schütze, H. 2008. *Introduction to information retrieval*. New York: Cambridge University Press.

Maulik, U., & Bandyopadhyay, S. 2002. Performance evaluation of some clustering algorithms and validity indices. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, *24*(12), 1650-1654.

Miller, H. J. and Han, J. 2009. *Geographic data mining and knowledge discovery*. Boca Raton, FL: CRC Press.

Moran, P. A. 1950. Notes on continuous stochastic phenomena. *Biometrika*, 37 (1/2), pp. 17-23.

Nazemi, A. R., Akbarzadeh-T, M. R., & Hosseini, S. M. 2003. Fuzzy systems as a fusion framework for describing nonlinear flow in porous media. In proceedings of *Fuzzy Information Processing Society, 2003. NAFIPS 2003. 22nd International Conference of the North American*: IEEE. pp. 389-394.

Nöllenburg, M. 2007. Geographic visualization. In *Human-Centered Visualization Environments* (pp. 257-294). Springer Berlin Heidelberg.

O'sullivan, D. and Unwin, D. 2003. *Geographic information analysis*. Hoboken, N.J.: Wiley.

Piatetsky-Shapiro, G. and Frawley, W. 1991. *Knowledge discovery in databases*. Menlo Park, Calif.: AAAI Press.

Rantanen, H., & Kahila, M. 2009. The SoftGIS approach to local knowledge. *Journal of environmental management*, *90*(6), pp. 1981-1990.

Rousseeuw, P. J. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20 pp. 53-65.

Ruspini, E. H. 1970. Numerical methods for fuzzy clustering. *Information Sciences*, *2*(3), 319-350.

Shneiderman, S. B., & Plaisant, C. 2005. Designing the user interface 4th edition. *ed: Pearson Addison Wesley, USA*.

Silverman, B.W. 1986. *Density Estimation for Statistics and Data Analysis*. New York: Chapman and Hall.

Slocum, T. A., 2005. *Thematic cartography and geographic visualization*. Upper Saddle River, NJ: Pearson/Prentice Hall.

Špatenková, O. 2009. *Discovering Spatio-Temporal Relationships: A Case Study of Risk Modelling of Domestic Fires*. Ph.D. Helsinki University of Technology.

Špatenková, O., Demšar, U., & Krisp, J. M. 2007. Self-organising maps for exploration of spatio-temporal emergency response data. In *Proceedings of Geocomputation* (Vol. 2007).

Tan, P., Steinbach, M. and Kumar, V. 2005. *Introduction to data mining*. Boston: Pearson Addison Wesley.

Verd, J. M. and Porcel, S. 2012. An Application of Qualitative Geographic Information Systems (GIS) in the Field of Urban Sociology Using ATLAS. ti: *Uses and Reflections*. 13(2).

Zadeh, L. A. 1965. Fuzzy sets. *Information and control*, *8*(3), pp. 338-353.

Zimmermann, H. 1975. Description and optimization of fuzzy systems. *International Journal of General System*, 2(1), pp. 209-215.

Appendix A

Matlab fcm clustering output transformation:

```
[center,U,obj_fcn] = fcm(pos,7);
maxU = max(U);
index1 = find(U(1,:) == maxU);
index2 = find(U(2,:) == maxU);
index3 = find(U(3,:) == maxU);
index4 = find(U(4,:) == maxU);
index5 = find(U(5,:) == maxU);
index6 = find(U(6,:) == maxU);
index7 = find(U(7,:) == maxU);
figure
line(negs(index1,1),negs(index1,2),'linestyle','none','marker', 'o','color','k');
line(negs(index2,1),negs(index2,2),'linestyle','none','marker', 'o','color','r');
line(negs(index3,1),negs(index3,2),'linestyle','none','marker', 'o','color','b');
line(negs(index4,1),negs(index4,2),'linestyle','none','marker', 'o','color','y');
line(negs(index5,1),negs(index5,2),'linestyle','none','marker', 'o','color','w');
line(negs(index6,1),negs(index6,2),'linestyle','none','marker', 'o','color','c');
line(negs(index7,1),negs(index7,2),'linestyle','none','marker', 'o','color','g');
hold on
plot(center(1,1),center(1,2),'kx','markersize',10,'LineWidth',2)
plot(center(2,1),center(2,2),'kx','markersize',10,'LineWidth',2)
plot(center(3,1),center(3,2),'kx','markersize',10,'LineWidth',2)
plot(center(4,1),center(4,2),'kx','markersize',10,'LineWidth',2)
plot(center(5,1),center(5,2),'kx','markersize',10,'LineWidth',2)
plot(center(6,1),center(6,2),'kx','markersize',10,'LineWidth',2)
plot(center(7,1),center(7,2),'kx','markersize',10,'LineWidth',2)

for i=1:n
if ismember(i,index1)==1
IDXfcm(i,1)=1;
elseif ismember(i,index2)==1
IDXfcm(i,1)=2;
elseif ismember(i,index3)==1
IDXfcm(i,1)=3;
elseif ismember(i,index4)==1
IDXfcm(i,1)=4;
elseif ismember(i,index5)==1
IDXfcm(i,1)=5;
elseif ismember(i,index6)==1
IDXfcm(i,1)=6;
elseif ismember(i,index7)==1
IDXfcm(i,1)=7;
end
i=i+1;
end
```

Appendix B

Python code for creating binary transactional database:

```
n= sum(1 for line in open('file.txt'))
i=1
l=0
lis=[]
certain=set([0])
result=open('result.txt','w')
for line in open('file.txt').readlines():
      s2=set([line.strip()])
     certain=certain.union(s2)
while i<=n:
   x=str(i) in certain
   #print(x)
   if x==True:
      result.write('1 \n')
   elif x==False:
      result.write('0 \n')
   i+=1
result.close()
print('done')
```