Journal of Geosciences and Geomatics, 2015, Vol. 3, No. 1, 1-6
Available online at http://pubs.sciepub.com/jgg/3/1/1
© Science and Education Publishing
DOI:10.12691/jgg-3-1-1

A Spatio-qualitative Knowledge Discovery Paradigm

Kamyar Hasanzadeh*

Department of Planning and Geoinformatics, Aalto University, Espoo, Finland
*Corresponding author: kamyar.hasanzadeh@aalto.fi

Received April 10, 2014; Revised January 10, 2015; Accepted February 12, 2015

Abstract 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. Similar to other types of geocoded data, various spatial, visual, analytical, and exploratory techniques can be applied to the spatio-qualitative datasets in order to discover knowledge. Typical spatial analysis provides techniques for discovering patterns from large geographical datasets. However, due to qualitative characteristics of this type of data, these techniques should be used more strategically in order to achieve concrete knowledge. Accordingly, this research propounds a four stage spatial knowledge discovery strategy that is adapted to meet the most common characteristics of the spatio-qualitative data specifications. Furthermore, the proposed paradigm is applied to a case study of urban impression in Helsinki region in Finland, and the results are briefly presented.

Keywords: *spatio-qualitative, SoftGIS, spatialdata analysis, knowledge discovery, GIS*

Cite This Article: Kamyar Hasanzadeh, “A Spatio-qualitative Knowledge Discovery Paradigm.” *Journal of Geosciences and Geomatics*, vol. 3, no. 1 (2015): 1-6. doi: 10.12691/jgg-3-1-1.

1. Introduction

Many fields of geographic research are observational rather than experimental, because the spatial scale is often too large and geographic problems are too complex for experimentation [13]. New knowledge is acquired by searching for patterns, formulating theories, and testing hypotheses with observations. With the continuing efforts by scientific projects, government agencies, and private sectors, huge geographic data have been, and continue to be, collected. We now can obtain much more diverse, dynamic, and detailed data than ever possible before with modern data collection techniques, such as global positioning systems (GPS), high-resolution remote sensing, location-aware services and surveys, and internet-based volunteered geographic information [2]. Generally speaking, geography and related spatial sciences have moved from a data-poor era to a data-rich era [14].

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. This understanding is the ultimate goal of a spatial knowledge discovery process. Knowledge discovery is an iterative process that involves multiple steps, including data selection, cleaning, preprocessing, and transformation; incorporation of prior knowledge; analysis with computational algorithms and/or visual approaches, interpretation and evaluation of the results; formulation or modification of hypotheses and theories; adjustment to data and analysis method; evaluation of result again; and so on [1]. Knowledge discovery naturally fits in the initial stage of a deductive discovery process, where researchers develop and modify theories based on the discovered information from observation data [14].

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 [20]. This has contributed to the considerable growth in acquisition of geocoded qualitative data or so-called spatio-qualitative data. Studying qualitative data in their geographical context has the potential to reveal useful information in different studies involved in these two types of information.

2. Spatio-qualitative Data

Spatio-qualitative data refers to a wide range of qualitative and subjective data that have been geographically located [4]. In other words, spatio-qualitative data is a collection of qualitative information which has been geocoded at the same time as being collected. There is a wide variety of such data varying from textual and non-textual narratives, through multimedia, to internet-based surveys. One of the dominant efforts in this context is the acquisition of SoftGIS data. SoftGIS, developed by MarkettaKyttä 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 [7,8]. 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 [8], [17].

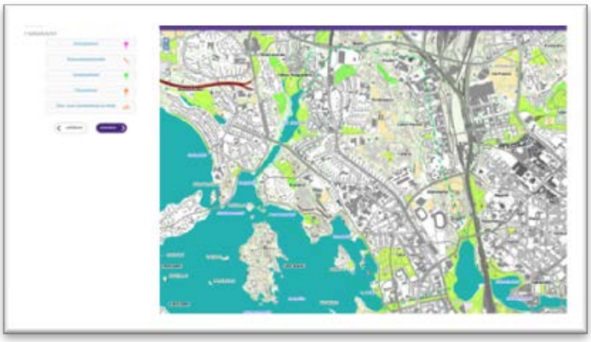

Figure 1. SoftGIS survey panel preview: http://demo.asiatkartalle.fi/ (accessed on 5.2.2014)

Spatio-qualitative datasets are different from other common types of spatial datasets in many senses.Spatio-qualitative datasets are a combination of spatial and qualitative information. These two types of information are of distinct natures and therefore require different processes in order to be studied and generate useful knowledge. Moreover, most of the acquired spatio-qualitative datasets consider each entry as a one-dimensional entity. For instance, in SoftGIS data each person’s local experience has been represented as point while in reality it may be associated with a feature of any shape, size, and dimensions (a road, a rectangle park, or a round square). Similarly, in studies working with narratives, each narrative has been assigned to a geographical coordinate that represents a point rather than an area. Furthermore, qualitative materials are the direct representatives of the human thoughts, experiences and even emotions. Human experiences are not necessarily homogenous and they can considerably vary from a person to another. These and many other distinguishing characteristics [4] form a realm of information which is filled by ambiguities and various types of uncertainties. This highlights the crucial role of a structured strategy that can organize one’s approach in studying such data.

3. Four Stage Spatio-qualitative Knowledge Discovery

The discussed characteristics of the spatio-qualitative data highlight the considerable need for an organized approach towards studying these datasets. This research is motivated by this needs and attempts to fulfil it by proposing a four stage spatio-qualitative analysis process that can achieve an inclusive understanding of the data while studying it from different aspects. This paradigm consists of four steps: *data processing, data analysis, data visualization, and spatial data mining*. A schema of the paradigm can be observed in Figure 2.

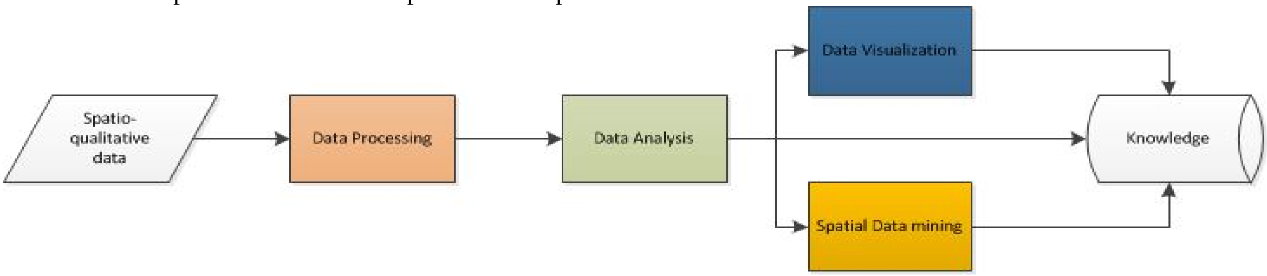

Figure 2. Spatio-qualitative knowledge discovery paradigm

3.1. Data Processing

Data processing is an inseparable part of any data mining and knowledge discovery process that deals with large amounts of data. A large dataset may have various types of anomalies that need to be spotted and fixed in order to provide the required basis for a concrete and reliable analysis. Moreover, depending on the type of analysis that one is going to perform on a dataset, one may require certain structures of dataset with appropriate classes, formats etc. In other words, *data processing* is a series of tasks which are intended to prepare the dataset by data cleaning, classification, data restructuring, etc. Moreover, while taking a close look at the dataset, this is a good opportunity to hypothesize about the possible sources of uncertainties that need to be further tested in next stages.

3.2. Data Analysis

Spatio-qualitative data usually reflect a large number of people’s perceptions and experiences of spatial locations [8]. 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 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 [16].

Data analysis stage aims to identify the principal patterns and characteristics of the dataset through a range of statistical and computational methods. Particularly, through this stage, the spatio-qualitative dataset’s characteristics are explored and the findings are later taken into account in the other two stages. As it can be observed in Figure 2, spatial analysis not only provides the required foundation of other analytical stages, but also contributes to the knowledge discovery by identifying the gist of the information and patterns. Based on the aim of a study and the specific characteristics of the case study, a spatial analysis may consist of following sub-processes:

- Spatial distribution analysis
- Spatial segmentation analysis
- Pattern detection

- Directional effects
- Etc.

3.3. Data Visualization

Spatio-qualitative data deal with qualitative information. Accordingly, not all its underlying information can be perceived and described through verbal and mathematical descriptions. Moreover, the potential users of the study 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 [18]. Moreover, through visualization, it is typically easier for the user to understand [9]. 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 [19]. Maps are the primary tool to present, use, interpret, and understand spatial data [15]. 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 [15].

Visualization of the spatio-qualitative data is usually more challenging than visualizing most other common types of spatial data. That is because the data typically contains huge masses of points, most of which are overlapping. Moreover, existence of many contradicting records in a short distance from each other is common to this type of data. The reason is that the spatio-qualitative data represents people's personal experiences and thus it can significantly vary from one person to another. Furthermore, the spatio-qualitative data typically consists of more than three dimensions. Depiction of multi-dimensional datasets is one of the oldest challenges of cartographers and visual analysts.

3.4. Spatial Data Mining

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 kinds of knowledge that is not explicitly stored in dataset [10,11,12]. 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 [14]. Using these computational techniques can help spatial scientists to make more profound and concrete discoveries while studying geocoded and spatial datasets. Previous steps of this paradigm aimed to make discoveries through using different visual and analytical exploratory techniques. The purpose of this step is to reach a new, and perhaps more profound, level of knowledge discovery, by using more statistically supported spatial data mining techniques.

4. Case Study: SoftGIS and urban Impression in Helsinki region

SoftGIS is an innovative attempt in acquisition of geocoded (hard) qualitative (soft) data that is designed to generate spatio-qualitative datasets, with applications in urban planning, through map-based surveys. SoftGIS data reflects a large number of people's perceptions and experiences of spatial locations [8]. 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. Accordingly, using the paradigm presented in this article, we will study the SoftGIS data with and aim of discovering useful knowledge.

The following two datasets are used in this case study:

- The SoftGIS data was provided by Mapita Ltd. and it included people's recorded urban experiences in Espoo and Helisnki region, Finland.
- 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 [6].

4.1. Data Processing

The SoftGIS 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.

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:

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.

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. Moreover, considering the acquisition procedure of SoftGIS we can hypothesize that the data consists considerable cognitive uncertainty [14].

4.2. Data Analysis

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 [21].

In this study spatial statistics toolbox of ArcGIS is used to calculate the Moran's I. The results are presented in Table 1.

Table 1. 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 1) 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 [22].

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 [22]. 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.

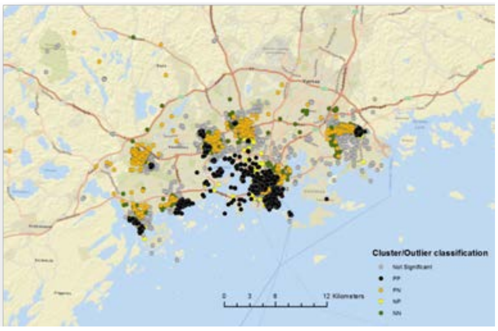

Figure 3. Anselin local Moran's I, Cluster/Outlier classification.

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 2.

Table 2. Correlation coeffcient values

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

The *r* values in Table 2 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 4).

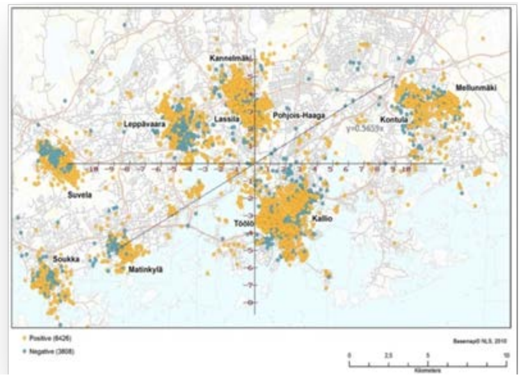

Figure 4. Correlation direction - The arrow represents the direction towards which the overall number of negative impressions rises

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.

4.3. Data Visualization

The SoftGIS urban impression map can be generated by calculating a weighted average of markings within a predefined cell [5]. 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 (Figure 5).

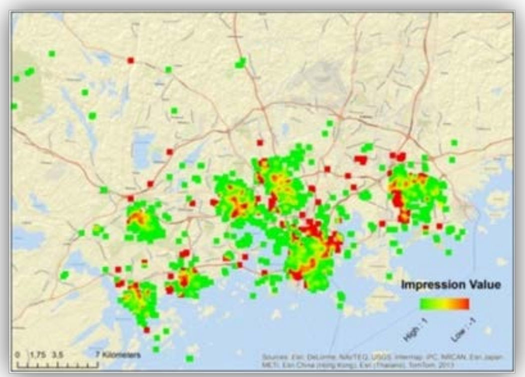
Impression Value
High : 1
Low : -1
0 1.75 3.5 7 Kilometers
Sources: Esri, DeLorme, NAVTEQ, USGS, Intermap, iPC, NRCAN, Esri Japan, METI, Esri China (Hong Kong), Esri (Thailand), TomTom, 2013

Figure 5. Weighted average impression map. From (Hasanzadeh, 2014b)

Studying this map provides useful information and contributes to knowledge discovery as described in [5].

4.4. Spatial Data Mining

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. Thus, a spatial data mining technique, namely spatial *Association rule mining* or co-location, is applied in order to discover more profound and statistically supported associations between the SoftGIS dataset and building types.

A summary of these association rules can be observed in Figure 6.

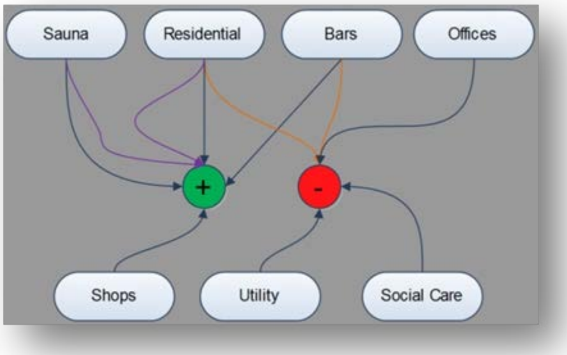

Figure 6. A summary of the discovered associations between building types and impressions. Black arrows represent singular antecedents and the other same-color arrows indicate pair antecedents

As it can be speculated in Figure 6, 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.

5. Conclusion

This study brings a broad viewpoint on the process of knowledge discovery within spatio-qualitative datasets. Spatio-qualitative datasets are typically complex and require a multi-aspect analysis approach in order to reveal useful knowledge. This study designed a paradigm that starts with data processing and proceeds towards knowledge discovery through spatial analysis, data visualization, and spatial data mining. The case study used in this study, took advantage of this paradigm and the discovered knowledge are a proof of this approach's potentials in spatio-qualitative knowledge discovery.

Apart from the characteristics that are common to the spatio-qualitative data, each type of them has also some unique characteristics. Therefore this paradigm may require adaptation to the case study and being tailored to meet its specifications. However, in a big picture, this research provided sample criteria of how the spatio-qualitative data should be treated and it presented a knowledge discovery paradigm that can be considered as a cornerstone of future studies.

Acknowledgement

This research is a part of the master's thesis conducted at Department of Planning and Geoinformatics in Aalto University, Finland. My deepest appreciations go to Professor KirsiVirrantaus for giving me the opportunity to work in the department of Geoinformatics lab. I would also like to thank Dr. JussiNikander and Dr. Paula Ahonen-Rainio for their brilliant comments and suggestions on my work.

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

References

[1] Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. From data mining to knowledge discovery in databases. *AI magazine*, 17 (3), p. 37.

[2] Goodchild, M. F. 2007. Citizens as sensors: the world of volunteered geography. *Geo Journal*, 69 (4), pp. 211-221.

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

[4] Hasanzadeh, K. 2014. *SoftGIS Data Mining and Analysis: A Case Study of Urban Impression in Helsinki*. Master's. Aalto University (a).

[5] Hasanzadeh, K. 2014. Spatio-Qualitative Data Visualization: SoftGIS and Weighted Average Visualization. *Journal of Geosciences*, 2 (1), pp. 38-41 (b).

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

[7] Kahila, M. and Kyttä, M. 2006. The Use of Web-based SoftGIS-method in the Urban Planning Practices.

[8] Kahila, M. and Kyttä, M. 2009. SoftGIS as a bridge-builder in collaborative urban planning. *Springer*, pp. 389-411.

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

[10] 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.

[11] 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*.

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

[13] Mennis, J. and Guo, D. 2009. Spatial data mining and geographic knowledge discovery—An introduction. *Computers, Environment and Urban Systems*, 33 (6), pp. 403-408.

[14] Miller, H. J. and Han, J. 2001. *Geographic data mining and knowledge discovery*. London: Taylor & Francis.

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

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

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

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

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

[20] 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).

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

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