

Development of business spatial analysis tools: methodology and framework.

Gautier DARAS^{a,b}, Bruno AGARD^{a,c}, Hadrien CAMBAZARD^{d,e}, Bernard PENZ^{d,f}

^a FORAC, Ecole Polytechnique de Montreal, Montreal, QC H3C 3A7Canada

^b e-mail: gautier.daras@polymtl.ca

^c e-mail: bruno.agard@polymtl.ca

^d Univ. Grenoble Alpes, G-SCOP, F-38000 Grenoble, France

CNRS, G-SCOP, F-38000 Grenoble, France

^e e-mail: hadrien.cambazard@grenoble-inp.fr

^f e-mail: bernard.penz@grenoble-inp.fr

Abstract: In the aim of minimizing time wasting and tedious tasks in spatial analysis tools development, this paper introduces a methodology and a framework. The five steps methodology is composed of data gathering, data management, analysis model development, user interface development and tools validation. The framework, based on client-server architecture, allows collaborative work and easy navigation through methodology steps.

Keywords: GIS, Geo visualization, data mining, KDD, open source, spatial data mining.

1. INTRODUCTION

Competition in retail distribution is constantly growing, forcing companies to improve their methods. There are different strategies to gain these improvements, each of which leads decision makers to make choices that can impact a company's future. To be able to make the best choices, decision makers require access to a complete and precise understanding of the reality. To fulfill that need, data analysis and data mining are widely used methodologies. Research in these fields continue to grow, especially with the development of new technologies.

One particular point about the retail sector is that the environment has a direct influence on the success of retail outlets (Cliquet, Fady, & Basset, 2006). Thus, it is advantageous for retail companies to identify the environmental factors that influence success.

A good way to understand environmental effects is to integrate geo-spatialized data analysis with Geographic Information Systems (GIS).

However, there is limited research that attempts to combine these two approaches. The combination must be done properly in order to give understandable and valuable knowledge to decision makers. The analysis must be based on strong and recognized data mining methodology, and the extracted information must be easily accessible and understandable.

Our research aims at combining geospatial analysis and visualisation approaches to develop a business analysis tool. Firstly, the tool will help the decision maker to understand the environmental effects. Secondly, it will provide automated support for decision making. To that end, a research methodology and a development platform have been

developed. This paper aims to introduce and explain them. After a state of the art, the methodology will be exposed. Then, a framework design will be proposed and finally, an operational implementation of this design will be presented.

2. STATE OF THE ART

This section will firstly introduced data mining and knowledge discovery from data (KDD) concepts. Secondly GIS will be explained. Thirdly, the integration between them will be presented.

2.1 Data mining

Data mining has been defined as "the efficient semi-automated discovery of non-trivial, implicit, previously unknown, potentially useful and understandable information from large data sets"(Anand & Buchner, 1998).

Apte (1997), defines it as a process by which accurate and previously unknown information is extracted from large volumes of data. This information should be in a form that can be understood, acted upon, and used for improving decision processes. Apte (1997) add that data mining techniques can be divided into three major categories: predictive modeling, clustering, and frequent pattern extraction.

Data mining is often used within the process of Knowledge Discovery from Data (KDD). This process was introduced by Fayyad, Piatetsky-Shapiro, and Smyth (1996), and consists in nine principal steps: the first step is to develop an understanding of the application's domain and the relevant required knowledge but also to identify the goal of the KDD process from the customer's viewpoint. The second, third, and fourth steps consist of data gathering, cleaning,

preprocessing, and reduction. Those three phases should make the data ready to pass through data-mining process.

This leads to the fifth, sixth, and seventh steps: finding a data mining model category that fits the customer's goal, finding the models to use within this category, and applying it to find a pattern of interest. The two final steps concern the interpretation of the mined patterns, and acting on the discovered knowledge.

2.2 Geographic Information System

Geographic Information Systems (GIS) are “automated systems for the capture, storage, retrieval, analysis, and display of spatial data”(Clarke, 1990).

Hess, Rubin, and West (2004) advance that unlike paper maps, GIS are capable of storing, manipulating, and displaying a much richer set of attribute information.

Spatialized data features can be represented in different shapes within GIS. For example, polygons could represent buildings, grey lines could be associated with roads, and points can represent bus stations. Spatialized data features can be grouped within layers.

A layers does not only contains the features shapes, it also contains all the information accessible from the feature data. For example, the length for the roads, the number of habitants of the buildings, and for the bus stations, the bus route numbers.

The greatest strength of GIS is that they allow users to find spatial relations among features in the same or different layers. Therefore, it is possible to create new pertinent data depending on our research interest.

It might be interesting to know the average road distance between successive bus stations of a particular bus line. For other research purposes it will be useful to know if every building is within buffers distance from bus station. A lot of other spatial operations could be done to create valuable data.

GIS can also be used as geo-visualization tools, MacEachren and Kraak (2001) advance that geo-visualization integrates approaches from scientific computing, information visualization, data exploration analysis and other. He adds that those gathered approaches should make spatialized data presentation easily accessible.

For Hernandez (2007), Geo-visualization aims to transform; to turn large heterogeneous data into information (interpreted data) and subsequently, into knowledge (understanding derived from information).

Geo-visualization techniques are geared to exploit visual-cognitive abilities: such as, pattern recognition, ordering, and interpretation of visual cues.

Moreover, Andrienko, Andrienko, and Gatalsky (2003) advance that, in comparison to paper maps, computer-based visualization tools have two principally new properties: interactivity and dynamics.

2.3 Integration of data mining and GIS

Integration of geo-spatialized data analysis and GIS can be done in two main ways (Le, 2009):

- Spatial data mining, which consists of the extraction of implicit knowledge and spatial relation from spatialized data (Koperski & Han, 1995). It can consist, for example, of finding relationships among spatialized elements such as service proximity or density presence.

- Geo-visualisation, that takes advantage of the fact that humans learn more easily and efficiently when the support is visual rather than when it is textual or numerical (Lloyd, 1997). Within GIS, an abundance of knowledge could be easily accessible to the user without overwhelming him. The ease of access to information could be done by allowing the user to focus on points of interest, or to get different visualisation formats.

As data mining results are generally incomprehensible for end users, results should be shown within user friendly interfaces, such as geo-visualization tools. Unfortunately, there is limited research that tries to combine spatial data mining and visualization. It follows that specific methodology and framework for this combination has not yet been set up.

In the next part, an extended version of Fayyad et al. (1996) methodology, will be introduced. This methodology incorporates aspects of dealing with spatialized data and geo-visualization.

3 METHODOLOGY

There are two main differences between the KDD process previously outlined (Fayyad et al., 1996) and our proposed methodology (figure 1). Firstly, in this process, spatial data have to be taken into account. Secondly, a user interface has to be developed to allow end users to easily understand the results of the KDD process.

The three first steps of the KDD process are grouped in a phase called knowledge building; data gathering and cleaning. After this preliminary phase the fourth step is changed into a phase called non spatial and spatial data treatment.

Non spatial and spatial data treatment

The non-spatial data treatment consists in the same tasks as the initial fourth step: finding useful features to represent the data depending on the goal of the task. The change is in the spatial data treatment.

As we have seen before, GIS could allow us to find many relations among spatialized features. However, this spatial relations research should be done carefully. Indeed, if we take the example in section 2.2, in which spatial relations among bus stops and buildings were searched, the nature of the searched relations has to be well specified. For example if the bus service quality seems important in the project, a variable that measure this service quality should be created.

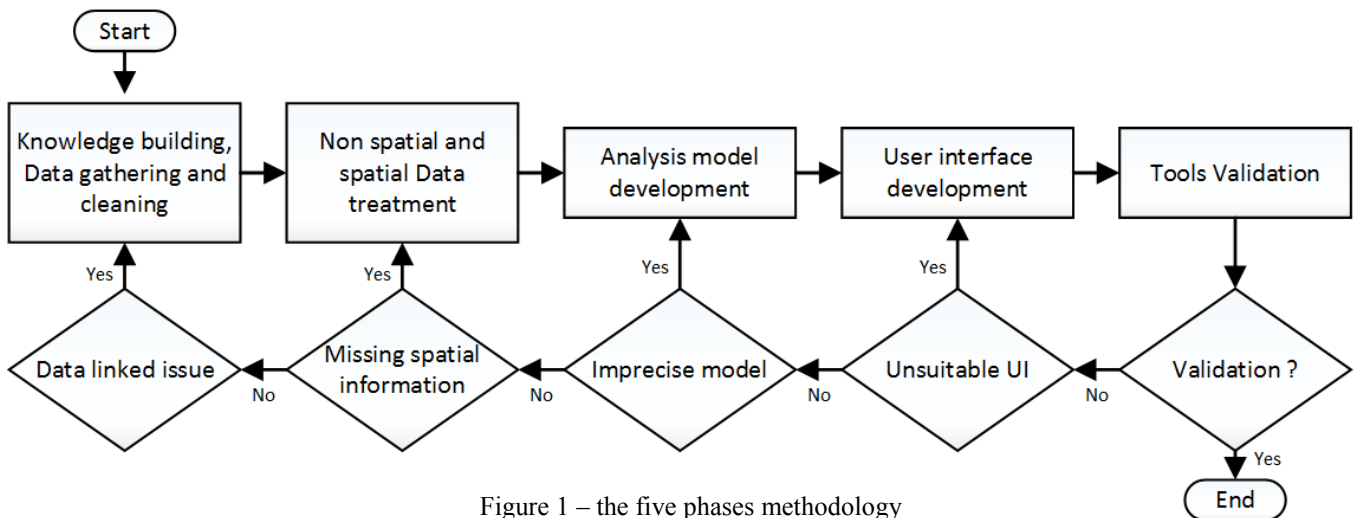


Figure 1 – the five phases methodology

The spatial relation that represents service quality could be computed in many ways. In the example case, do we want to have the average distance between bus stations? Or maybe the average distance among each buildings and its nearest bus station is more suitable information? And then, what do we call a distance here, crow flies distance? Or walking distance?

This example tries to explain why the preprocessing step from KDD process is transformed to a whole phase in our methodology. This phase will have to be done carefully and with specific application domain knowledge.

Then the steps 5 to 8 from KDD process are grouped into the analysis development model phases. The next difference is in the development of a user visualization interface.

User interface development and tool validation

As part of the project is to make the information easily understandable to decision makers, the user interface development step is added.

Hess et al. (2004) have advanced that the quality of tools increases with the competence of the decision maker. To be able to use this asset, the tools validation phase is added to complete our methodology. The two last phases replaces the ninth initial step.

For the UI development phases, the research should start by defining which information should be pertinent and accessible for the users. The visualization formats of this information should be wisely selected to maximize user understanding. An implementation of the user interface (UI) must then be done to ensure the usability of it.

The last phase, consisting of the tool validation, has to be made with the client/end user. In this step it will be ensured that the four precedent steps have been well realized. The UI must be understood, the options it provides must be used and the model should provide coherent and helpful results. Researchers should listen to all users' observations and identify which of the two new possible steps back it is about.

Unsuitable User Interface

The UI could be unusable for different reasons, the interface could be non-adapted to user computers, it could be latency issues, or readability problems. Whatever the difficulty is, it should make the researcher go back to the UI development phase.

Missing Spatial Information

In the case of spatial information lack, the researchers must return to phase one to get new pertinent data, or to phase two, in order to create new valuable indicators or spatial relation.

There are a lot of problems that could delay this kind of research project. Phase switches will happen often, as well as data transfer, and this could lead to time wasting. To avoid that, a framework has been designed and implemented in order to easily apply the methodology. This framework is presented in the following section.

4. FRAMEWORK PRESENTATION

4.1 Issues to be considered

The proposed methodology may contain some time wasting or tedious tasks. The next four sections will introduced some of the identified issues that can be encountered in this kind of project. For each identified issue, some tools that can address it will be introduced.

Working with adapted tools

In every research project, working with adapted tools is essential. In this particular project we identify four competence fields where adapted tools must be possessed: data storage, data treatment, data analysis, and data visualization. All four fields concern spatial and non-spatial data.

With non-adapted tools, some work should be done from scratch, while in adapted ones a lot of useful libraries exist to help the researchers in the project.

- For data storage a lot of robust database exist in the market, but only some of them allow spatial data management. One with an administration interface that allows faster treatment and administration should be used.
- To be able to work deeply with spatial data, the use of strong and robust GIS software is essential.
- The data analysis and data mining model should be made with strong and recognize software, to allow specific analysis within set parameters. One that allows spatial data analysis should be preferred.
- Good visualization tools should simplify the tasks of sorting through the different formats of the results of the analysis.

Communication among tools

Another issue is the communication among the tools. They should use the same data format, the researchers cannot allow the time wasting of data conversion tasks.

- Communication interfaces must be set up to among the different tools to automatize the data translation.

Work versioning and backup

A frequently encountered problem is the loss of work; it can be data loss, or the developed model not working anymore.

Not being able to access to previous versions of the model or to data backup will be a serious issue in that case.

- Version control systems have the ability to revert a document to a previous revision. They keep track of any modification of the document you want to version control, and for each version a backup is available. The document can be for example, a database dump or a project's source code.

UI design and implementation

UI development is not the strongest competence field of every researcher. Thus, it can be really time consuming for an unskilled developer to be able to set up a geo-visualization environment.

- A web map server, connected to the database, should give easy access to data visualisation through a web library.

Collaborative work

A lot of tasks and data could be redundant among researchers' project parts. Therefore, a collaborative work architecture that allows data or source code sharing and gathering could be a real time saver.

In the next section we introduce a framework based on an architecture that allow collaborative work and integrate previously presented tools.

4.2 Proposed framework design

Previous issues being considered, the following design is proposed.

To allow collaborative work, a client-server architecture has been chosen. In a classic fat server-clients approach the server could be overburdened by data access and geospatial analysis In the Rich-client Light server approach, the same

limitations could occur on client side (Vatsavai, Shekhar, Burk, & Lime, 2006).

To avoid these, our proposed architecture is composed by two types of clients (figure 2): researchers as rich clients and end users as light clients.

Thus researchers will proceed to the development and analysis of their own machine within robust software.

The end user will have access to visualisation tools through a basic web browser.

Rich clients will include GIS software, database administration tools, and analysis software.

Servers will host databases, version control system, and a web map server.

The end users just have to possess a non-obsolete web browser on his computer.

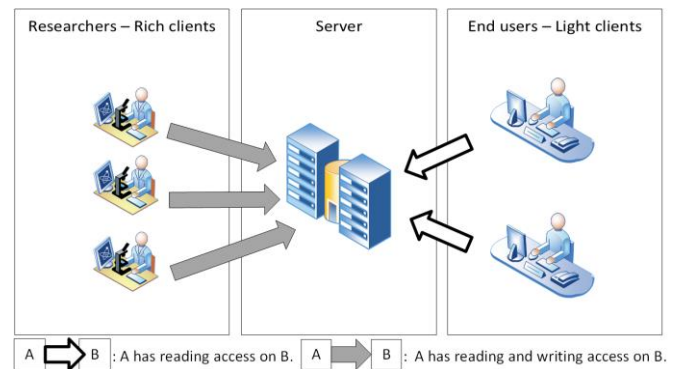


Figure 2 - the "rich client - server - light client" architecture

This framework design is the subject of an implementation in the following part.

4.3 Framework implementation

Technological needs are now presented, some software will be proposed to implement this design. Other software can be set up together as an implementation if they can communicate together. Thus we will describe the implementation choices.

The implementation of the designed framework has been led by the idea of building it over open source technologies.

Beside that it allows a relatively low cost implementation, open source technologies are used by a lot of researchers allowing for the creating and sharing of new software and knowledge. The use of open source technologies is the promise of new ways to organize science and heighten the pace of knowledge discovery (Rey, 2009).

In the following section we quickly introduce the open source software that composes the framework, and in a final part the global design is presented and explained.

The framework implemented with open source softwares

Most of the tools presented here are chosen within the Boundless suite, formerly OpenGeo Suite . The interaction between those tools is made easier as they are part of the same collection. The following are the tools from this suite:

- PostGIS is an extension that turns the PostgreSQL database system into a spatial database.

- GeoServer allows users to share and edit geospatial data. It can publish data from any major spatial data source using open standards.

- OpenLayers is a JavaScript library for building mapping applications in a browser.

The other used software are the following:

- QGIS is a well-known GIS software interacting easily well with PostGIS database.

- R is a free software environment for statistical computing and graphics. It is widely used and recognized, and has extensions for spatial analysis.

- PgAdmin is an administration interface of PostgreSQL database than can deal with PostGIS database.

- Git is a widely used distribution revision control system.

Final implementation

The final implementation is represented in figure 4. There are four main blocks, which have the following features:

The git repository is the place where the work of researchers, the algorithms, and the source code are placed. In this repository any modification is saved, and the code can be shared and reviewed by collaborators.

Every researcher of the team should install the following application on his computer: PgAdmin, R studio, QGIS in order to deal with data analysis and management.

The server is composed of two main parts: the PostGreSQL/PostGIS database container and the web map server Geoserver:

- The PostGreSQL/PostGis Database container can contain database for every researcher, and a common one for data

that could be accessible, but not modifiable, by every team member.

- The Geoserver should be able to give end users easy access to geographical data. This accessible data have to be easily interpretable within a web browser through the use of a web mapping library such as OpenLayers.

4.4 Discussions / Comments

Currently, different softwares are combined in order to easily navigate through the methodology phases. The implemented framework already allows to easily import different spatialized data format with PostGIS. It is then possible to treat those data sets within PgAdmin for cleaning or preprocessing. QGIS is then necessary to practice spatial data treatment. Finally Geoserver can easily render spatial information on a web browser, by allowing map services such as OpenLayers to access database allowed content (figure 3).

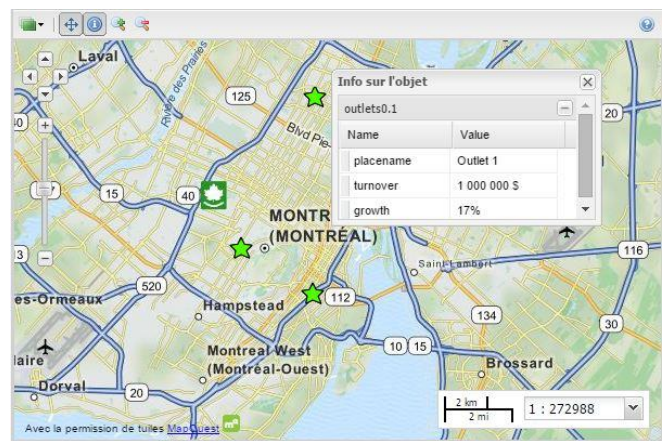


Figure 3 - Outlet information visualization

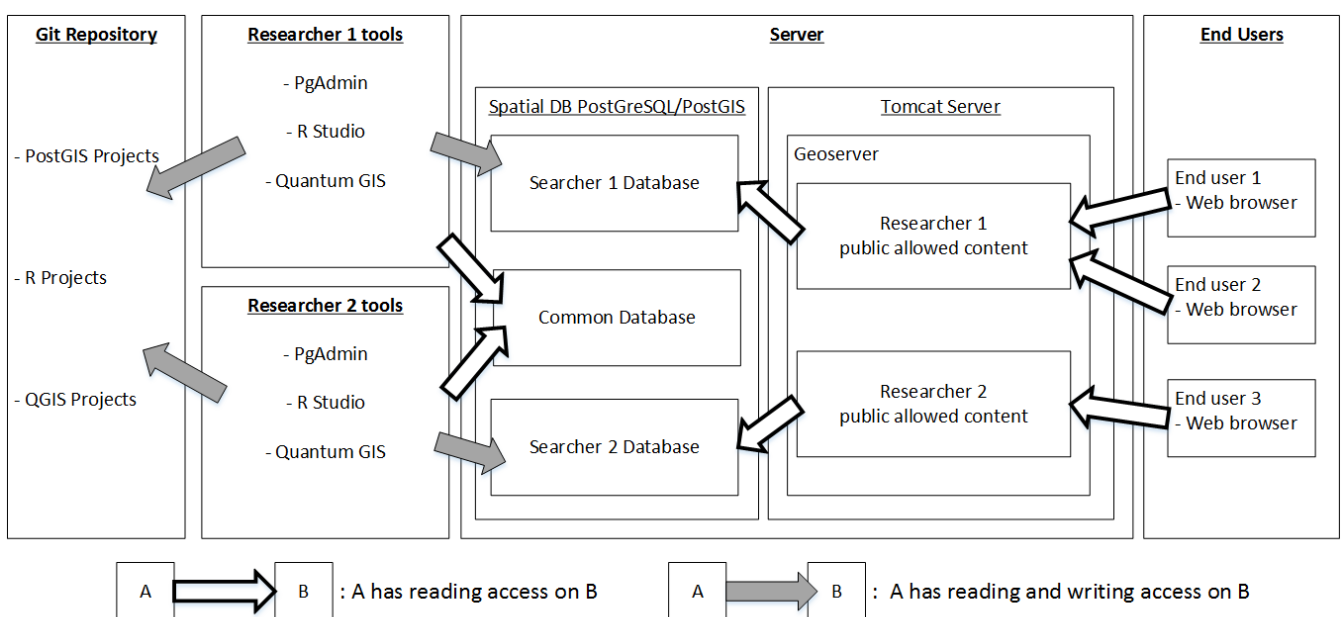


Figure 4 - Framework implementation with open source software

As the computation of model analysis is made on researcher's computers, the model complexity should accord with the computation capacity of the computers.

All the available softwares for each need were not tested, these tests could be useful to ensure the optimisation of the framework implementation.

The data management portion is more developed in this framework than the data visualization part. Future research regarding developed advanced user interface, would require knowledge on web based technologies (javascript, css, html).

No data mining techniques are implemented for now, the tools set up has been made to allow the application of data mining models. Those models will have to be wisely chosen, and well adapted to the domain application.

Finally, the actual tool does not integrate decision support systems. In the future, when environmental factors of success will be understood, the elements to optimize could be identified. Then operation research methodologies could be adapted and integrated within the system.

5. CONCLUSIONS

There were a lack of methodology and tools for integrating spatial data mining and geo-visualization. This integration has been first proposed with the aim of developing spatial analysis tools.

To ensure the proper conduct of this development, a well known KDD process has been adapted into a five phase methodology.

In order to facilitate the navigation through the methodology phases, a set of open source software have been put together, resulting in a development framework.

The feedback of real applications is now needed to refine the proposed methodology and framework.

REFERENCES

- Anand, S. S., & Buchner, A. G. (1998). Decision Support Using Data Mining. *Financial Times*.
- Andrienko, N., Andrienko, G., & Gatalysky, P. (2003). Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6), 503-541. doi: 10.1016/s1045-926x(03)00046-6
- Apte, C. (1997). Data mining: An industrial research perspective. *Ieee Computational Science & Engineering*, 4(2), 6-9. doi: Doi 10.1109/99.609825
- Boundless Geo suite website. Retrieved December 20, 2014, from <http://boundlessgeo.com/>
- Clarke, K. C. (1990). *Analytical and computer cartography* (Vol. 290): Prentice Hall Englewood Cliffs (NJ).
- Cliquet, G., Fady, A., & Basset, G. (2006). *Management de la distribution* (2nde ed.).
- Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *Ai Magazine*, 17(3), 37-54.
- GeoServer website. Retrieved December 20, 2014, from <http://geoserver.org/>
- Git website. Retrieved December 20, 2014, from <http://git-scm.com/>
- Hernandez, T. (2007). Enhancing retail location decision support: The development and application of geovisualization. *Journal of Retailing and Consumer Services*, 14(4), 249-258. doi: 10.1016/j.jretconser.2006.07.006
- Hess, R. L., Rubin, R. S., & West, L. A. (2004). Geographic information systems as a marketing information system technology. *Decision Support Systems*, 38(2), 197-212. doi: DOI 10.1016/s0167-9236(03)00102-7
- Koperski, K., & Han, J. (1995). Discovery of spatial association rules in geographic information databases. *951*, 47-66. doi: 10.1007/3-540-60159-7_4
- Le, T. T. H. (2009). *Perspectives d'integration entre le data mining et les SIG*. (M.Sc.), École Polytechnique de Montréal.
- Lloyd. (1997). *Spatial Cognition, Geographic Environments*. Dordrecht: Kluwer Academic Publishers.
- MacEachren, A. M., & Kraak, M.-J. (2001). Research Challenges in Geovisualization. *Cartography and Geographic Information Science*, 28(1), 3-12. doi: 10.1559/152304001782173970
- Openlayers website. Retrieved December 20, 2014, from <http://openlayers.org/>
- PgAdmin website. Retrieved December 20, 2014, from <http://www.pgadmin.org/>
- PostGIS website. Retrieved December 20, 2014, from <http://postgis.net/>
- QGIS website. Retrieved December 20, 2014, from <http://www.qgis.org/en/site/>
- R Project website. Retrieved Decemeber 20, 2014, from <http://www.r-project.org/>
- Rey, S. J. (2009). Show me the code: spatial analysis and open source. *Journal of Geographical Systems*, 11(2), 191-207. doi: 10.1007/s10109-009-0086-8
- Vatsavai, R. R., Shekhar, S., Burk, T. E., & Lime, S. (2006). UMN-MapServer: A High-Performance, Interoperable, and Open Source Web Mapping and Geo-spatial Analysis System. *4197*, 400-417. doi: 10.1007/11863939_26