

Mining spatial data repositories: lessons learned

J Pretorius¹, J Schoeman¹, HL Viktor², and C Blyth¹

¹*Department of Informatics, University of Pretoria, South Africa.*

²*School of IT and Engineering, University of Ottawa, Canada.*

Abstract

Spatial data mining refers to the extraction of implicit, nontrivial and previously unknown knowledge, in the form of spatial and non-spatial patterns, from spatial data repositories. These repositories include a number space-related data, such as maps, pre-processed remote sensing or medical imaging data, amongst others. In addition, the data repositories are usually ordered using complex spatial indexing structures.

Spatial data mining is extremely resource intensive and complex, since it usually involves the mining of huge amounts of spatial data from various diverse applications. The combination of diverse spatial data repositories into an integrated, subject oriented data warehouse aids the data mining practitioner to consolidate various data sources, thus potentially leading to a high quality repository to be used for further exploration.

This paper describes the lessons learned during a data mining research effort concerning a census spatial data warehouse. The census data repository concerned the 1996 census of the Greater Pretoria Metropolitan Council area located in the Gauteng Province of South Africa. Approximately 3 million people, residing in formal or informal settlement within rural and urban areas, inhabit this region. The aim of this research project, which was completed in two steps, was to investigate how spatial data mining may be used to discover new knowledge.. The lessons, as learned through the application of classification and association rule mining, are discussed.

1 Introduction

The knowledge era is characterised by the accumulation of large amounts of data. Since the 1980s, all major organisations have acquired large databases in which to store all the gathered information (Adriaans [1]). Although these databases had traditionally been used for operational purposes only, it is increasingly considered as a wealth of untapped information on which decisions may be based.

Knowledge Discovery from Data (KDD) describes the extraction of this potentially valuable knowledge from huge data repositories (Adriaans [1]). This extraction is normally done with automated data analysis techniques. The KDD process consists of five non-linear phases, namely data selection sampling, pre-processing, transformation, data mining and evaluation (MacEachren [2]). The actual knowledge discovery phase is called data mining. The data mining tool constructs a model of the data, which potentially provide new insights into the information as contained in the underlying data (Victor [3]). Due to the rapid progress of data collection tools, advanced database technologies and the World Wide Web technologies, data appears in increasingly complex forms. This includes structured data, unstructured data, hypertext and multimedia. As a result, research focuses increasingly on the exploratory data analysis and describes the lessons learned during a data mining research effort concerning census spatial data database.

An example of such complex data types, is spatial data. Due to the fact that large amounts of data has been accumulated in geographical information system (GIS) databases during the last few decades, more than enough possibilities for exploratory data analysis exists for using KDD methods and associated tools. In addition, it has become important to present processed information in a coherent and easily accessible way.

This paper discusses lessons learned during a data mining research effort concerning a Census spatial data warehouse. The paper is organised as follows. Section 2 introduces the concept of spatial data mining. This is followed by a description of the Census case study in section 3 after which section 4 discusses how the data mining of census data was done and the results achieved by the methods used. In Section 5, the lessons, as learned through the application of classification and association rule mining, are discussed. Finally, Section 6 concludes the paper.

2 Spatial Data mining

This section describes the difference between standard data mining and spatial data mining as well as the influence of the data on the way data mining tools are used.

The objectives of spatial data mining do not differ from the objectives of data mining in general. Both types of data mining are exploring data for certain patterns

and relationships in a hope of finding interesting and previously undiscovered information. Spatial data mining (SDM), or knowledge discovery in spatial databases, refers to the extraction of implicit knowledge, spatial relations, or other patterns not explicitly stored in spatial databases (Han [4]). A spatial database stores spatial data types as well as spatial relationships between these data types. Spatial data carries topological and/or distance information and it is often organized by spatial indexing structures. This data is then accessed by spatial access methods. SDM demands an integration of data mining technologies with spatial database technologies. It consists of three broad classes of data types, geographical coordinates, temporal coordinates, and multivariate attributes relating to the geographical entities. However, the application of conventional data mining tools on spatial data may lead to disregarding the special characteristics of spatial data and geographical analysis. According to Koperski [5], these distinct features of spatial databases pose challenges and bring new opportunities for mining information from spatial data. Statistical spatial analysis methods, normally associated with GIS, are currently the way of dealing with spatial data. However, these methods are restricted to user defined queries and simple analytical tasks. The questions arises thus on how data mining may aid the knowledge discovery process in spatial data (Han [4])

The next section describes the census case study where data mining was used in an attempt to discover knowledge in a spatial data repository.

3 The Census Case Study

The Census data repository concerned the 1996 Census of the Greater Pretoria Metropolitan Council (GPMC) area located in the Gauteng Province of South Africa. Approximately 3 million people, residing in formal or informal settlement within rural and urban areas, inhabit this area.

The total dataset has about 23 tables describing the amount of people in a certain geographical region that comply with the criteria given. Table 1 gives an example of some of the tables that are present in the dataset.

Table 1: Example of tables in GPMC 1996 Census database.

Employment status by region by population group and gender (15-65 years)
Region by fuel used for heating by population group of head of house and gender
Region by fuel used for lighting by population group of head of house and gender
Region and country of birth by gender by population group
Region and Population group by gender by Highest Education level
Region by Age by Population group and Gender
Region by Citizenship by Population group and Gender
Region by Disability summary by Population group and Gender
Region by First Language by Population group and Gender
Region by Individual Income (Monthly) by Population and Gender
Region by Industry by Population and Gender for Person weighted

Originally, this census repository was designed such that GIS software may be applied to it.. A map reference attribute (GIS Key) is always present in spatial data and it denotes a certain polygon or area on a map. For the GPMC data repository, 2606 regions exist for the greater Pretoria area. For each of the different regions, demographic information regarding the individuals was stored. For example, the age, citizenship, first language, race and gender, amongst others, are kept. The tables included information regarding the four major population groups (Black, Coloured, White and Asian-Indian) together with the Gender.

By using conventional data mining methods on spatial data, this research hoped to find additional relationships and trends and thus the possible discovery of new knowledge. The data also lends itself to the analysis of the quality of life in the different regions in and around Pretoria

The research started out with an exploratory analysis to determine the quality of data, as well as to see what the data say about the quality of life in and around Pretoria. For the purpose of this research, only attributes that could possibly have an effect on standards of living were selected e.g. income, cooking fuel, education level and water supply, lighting fuel etc.

For the actual data-mining phase, the researches used a suite, called RuleQuest (www.rulequest.com) to help them extract rules from the data. This suite contains three tools, Cubist, See5.0 and Magnum Opus. See5.0 analyses data to produce decision trees and/or rule sets that relate a case's class to the values of its attributes. Magnum Opus differs from See5.0 in that it finds association rules from data. This type of rules reveals underlying interactions between factors within the data. These rules are often used to gain insight into the drivers of behaviour, such as which products, when purchased, drive the purchase of which other products. Rule induction is normally used in datasets with either fields of high cardinality (many different values) or many columns of binary fields This program finds association rules from both basket data and attribute-value data. The last application, Cubist, analyses data and generates rule-based linear models, each of which is associated with a linear expression for computing a target value.

The next paragraph describes the application of these tools to the census data.

4 Data Mining of Census Data

This section describes the data mining experiments conducted on the selected subset of the census data repository.

4.1 Experiment 1: Quality of Life

The research was initially intended as an exploratory analysis to determine the quality of the data, as well as to see what the data reveals about the quality of life

in the Pretoria area. Cubist was the first tool to be used in the data mining research. This tool is normally used to generate rule-based predictive models from data.

Unfortunately, since Cubist is meant for continuous, numeric data, its application to the categorical and discrete census data is inappropriate. Since the number of people that fit in a particular profile was the target attribute, Cubist attempted to build numeric models regarding an average of the number of people that suited several profiles. Six rules were obtained using Cubist, but they were mostly of a trivial nature. It became clear that Cubist could be used only to get rules from numeric attributes. This made it almost impossible to get characteristics of a certain area. Furthermore, the rules that were gathered were difficult to analyse. Since the census data and Cubist are not compatible, it was decided to discontinue evaluation with this particular application.

The researchers then applied See5.0 on the dataset in an attempt to find better results. At first See5.0 was used with the “number of people” field as the target categorical attribute. This resulted in over 600 rules, many of which proved to be quite trivial. Some of the results, however, were quite interesting and valuable. The result pertaining to profiles that fit no people (i.e. a zero value for the “number of people”) were quite enlightening. For example, from the following rule it may be deduced that there are no white people, out of 206 database profiles, in the GIS Region 7030207.

Rule 16: (206, lift 1.2) GIS Key = 7030207 Pop Group Description = White -> class 0 [0.995]
--

Similarly, from the following rule, it can be inferred that three out of 840 observations are for homeless people.

Rule 6: (840/3, lift 1.2) Dwelling Description = None/homeless -> class 0 [0.995]

But unfortunately, just as with Cubist, many of the results yielded were not as significant as desired. The reason for this is that since See5.0 attempts to find patterns within categorical data, it attempted to find patterns with regard to specific number values. Therefore, since there are up to 308 different counts for the “number of people” that fit the profiles within the database, See5.0 constructed rules based on 308 different classes.

With this in mind, it was decided to categorise the “number of people” field in a new field (Number of People Category), and to use this new field as the target attribute. This data gave improved results, referring to one of the eight numbers of

people categories, rather than to individual numbers. The data could now furnish accurate results, although one should be careful in the interpretation thereof. For example, from the following rule, it can be deduced that three of the 840 observations of “coal” as the cooking fuel description were not for zero people. When interpreting the rule one has to be careful, since the three observations that do not fit this rule does not refer to three people, but three profiles that could each have any number above zero for the number of people.

Rule 9: (840/3, lift 1.2)
Cooking Fuel Description = Coal
-> class 0 [0.995]

Because of the difficulty with regard to the interpretation of the rules when the “Number of People Category” is the target attribute, it was realized that when making the target attribute one of the other categorical fields within the database, more meaningful results would be obtained. Therefore, using target attributes such as population group, gender and education level, created several different and more meaningful models.

In conclusion, See5 proved to be well suited to the census data, since it is meant for categorical data. Although several changes had to be made to the data, it proved to be quite useful in knowledge discovery. By changing the focus of the evaluation, i.e. the target variable, several different possible outputs could be created. The only problem that remains is the analysis of the amount of outputs and rules in each output. It was realized that it would be more logical to split this particular database into the different descriptive categories (e.g. education level, cooking fuel description, etc.) and do a separate analysis on each, depending on what area knowledge one want to study.

4.2 Experiment 2: Association Rule Mining

From the previous research it was learned that the data did not lend itself to the discovery of the desired type of knowledge. In order to find more useful rules, association analysis were tested after realizing that association rules show attribute-value conditions that occur frequently together in a given set of data [Han & Koperski, 2001].

In order to explore the quality of life in the Pretoria area, one must find associations between certain areas to build up a model of the characteristics of areas with regard to the populations that inhabit it. To be able to do this Magnum Opus had the best prospect of extracting differences in population group, gender and criteria.

During this set of experiments, additional data pre-processing transformed the data into a more appropriate format. Data pre-processing involved the transformation of the data to a different format. Magnus Opus was first used on the “number of

people category” attribute. After analysing the preliminary output, it was then realized that the tool attempts to find associations between factors in the database by “viewing” each row as an instance of the profile. Magnum Opus does not attach different weights to the profiles according to how many people are in the profile, and would therefore regard a profile describing 20 people with the same weight as a profile describing no people. Due to this, it was decided to expand the database in order to let each row represent one person.

4.2.1 Results obtained

The expanded database produced more accurate association rules, although the association rules still needed to be interpreted with caution. The effect of the nature of the data in the database on the quality of rules was again illustrated. For example, the interpretation of rule 1 and 2 given below: Association Rule 1 can safely be interpreted as: if a person’s education level is “Matric Only” then there is a 53.8 % (strength) chance that the person’s gender is female.

Association Rule 1:

Education Level=Matric Only -> Gender
Description=Female [Coverage=0.067 (12102);
 Support=0.036 (6509); Strength=0.538; Lift=1.34;
 Leverage=0.0091 (1641)]

Association Rule 2, however, cannot necessarily be interpreted as: when a person is female there is an 8.9% (strength) chance that their education level is “Matric Only”.

Association Rule 2:

Gender Description=Female -> Education Level=Matric
Only [Coverage=0.402 (72749); Support=0.036 (6509);
 Strength=0.089; Lift=1.34; Leverage=0.0091 (1641)]

Due to the nature of the database, misinterpretation is possible. As described above, each database row contains only one value for the attributes other than GIS Key, Gender and Population Group. As shown in the above rule, it is important to realize that Magnum Opus calculates that 40.2% (coverage) of the rows, or profiles, in the database describe females. However, not all of these profiles contain values for the “Education Level” field. Therefore, the calculations with regard to the “Education Level” are not accurate since every row in the database contains a value for Gender, but only a fraction of the database contains values for Education Level.

732 *Data Mining III*

Because of the nature of the database, Magnum Opus cannot find association rules between the variables that contain null values in certain parts of the database. Therefore, analysis that is more accurate can be conducted when splitting the database into various sections. For example, when a data file containing only the rows that have values for the “Education Level” field is evaluated, the following results are obtained:

Association Rules 3 & 4:

Education Level=Matric Only -> Gender Description=Female [Coverage=0.281 (12102); Support=0.151 (6509); Strength=0.538; Lift=1.03; Leverage=0.0046 (198)] Gender Description=Female -> Education Level=Matric Only [Coverage=0.521 (22440); Support=0.151 (6509); Strength=0.290; Lift=1.03; Leverage=0.0046 (198)]
--

It is clear that association rule 3 is the same as association rule 1. However, association rule 4 is now more accurate than association rule 2, since only the profiles containing values for the “Education Level” field are analysed. From association rule 4 it can now accurately be established that, of the females in the Education Level Data file, 29% have “Matric Only” as their highest education level.

5 Lessons Learned

It became clear in the research that knowledge is distinct from data, in that it can be seen as the meaningful product of the relationship between two or more objects of data. Decisions are made based on the extracted knowledge. In this research, the Rulequest data mining suite was used to search for patterns and relationships in the spatial data set.

Unfortunately, the rules extracted from the dataset did not reveal any useful patterns. Although many attempts were being made to change the data to a format compatible for the above mentioned data mining tools, no progress was made in uncovering interesting and useable rules. However, valuable lessons were learned.

Not all data are suited for data mining. According to Openshaw et al. [6], many data miners appear to believe that data mining tools will work with all data regardless of its subject origins. According to them, this merely perpetuates a myth that data mining is universally useful. It is also important to consider the special characteristics of the data you are working with. For example, spatial data mining is to be regarded as a type of data mining, similar to conventional data mining tools, but modified to take into account the special features of geo-information, the rather different styles and needs of analysis and modelling relevant to the world of GIS, and the peculiar nature of geographical explanation.

It can be argued that conventional data mining tools are still useful, in the same way that conventional statistical methods can be applied to spatial data even if they are not always adequate in the analysis and modelling of the data. Indeed, there is no doubt that conventional statistical tools continue to be useful as proven by the successes in industry. However, the argument presented here is that many of these general purpose data mining tools are not sufficiently focused on and tailored to the special needs of geographical analysis.

According to Openshaw et al. [6], some geo-informational data mining tasks may be successfully performed by conventional data mining software. For example, data reduction tools, such as multivariate classification, can be useful as a means of summarising the essential features of large spatial data sets (e.g. to create geo-demographics classifications). Similarly, modelling tools such as neural networks and decision trees can be readily applied to some geographic problems. It may be argued that while these methods ignore all of the special features of spatial data, they still "work" to some degree.

If one uses conventional data mining tools, one is forced to ignore the special characteristics of geo-referenced data. These data mining packages treat the X, Y co-ordinates (GIS Key variables) as ordinary variables (such as age, income, or education level). This is the reason why no definite and usable rules could be extracted from the dataset. There is no mechanism for handling location and spatial aggregation. In addition, it ignores the spatial entities and spatial concepts such as region, distance or map topology. Conventional data mining tools may be very powerful but they continue the geographical neglect inherent in conventional statistical methods.

Openshaw [6] argues that new types of data mining tools must be developed that can handle the special nature of spatial information, and also capture the spirit and essence of geographical-ness that a GIS-minded data miner would expect to have available.

Another factor that plays an important role in finding useful rules is the quality of the data. This factor has according to Viktor [7], a substantial impact on the results of the KDD Process. The quality of the census dataset is well suited for the use of a GIS, but not for a conventional data mining tool. The type of rules that were discovered was not suitable. For this type of data to be more suitable for a data mining tool, more attributes and information need to be included. The information in the dataset is all about the amount of people in certain areas that fulfil a certain criteria about education, occupation etc. This is unfortunately not enough to find associations between areas.

Although no usable findings were found in the dataset, it does not mean that data mining is not suitable for geo-referenced data. Spatial data mining according to Han et al. (2001), have wide applications in GIS, geo-marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control,

environmental studies, and many other areas where spatial data are used. A crucial challenge to spatial data mining though, is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data and the complexity of spatial data types and spatial access methods.

Another lesson learned is that constant human intervention is needed to get the right information out of a data set. Data mining users should be made aware that data mining programs are not magic tools to get instant information and rules from a dataset. In order for a data-mining tool to come to its fullest potential, the user must have an active role in analysing the data and perhaps resubmitting it for further analysis. The way one prepares the data at the outset will have a profound affect on the information one will eventually find.

It is important to have predetermined goals with regard to what knowledge areas to investigate. Since many different outputs, each containing hundreds of rules, are possible, one should have a good idea of what to look for.

For the experiments described here, it has been observed that more valuable and correct results are possible when splitting the database up into the different attributes, such as education level, cooking fuel description, income level, toilet type, etc. Applying the Data Mining tools to the separated data files resulted in not only more accurate outputs, but also allowed the interpreter to focus on a particular knowledge area at a time. This makes the possibility of gaining a more in-depth understanding regarding a particular field, e.g. education level, more probable.

6 Conclusion

Regardless of not finding any really usable rules from the dataset, the exercise was well worth the trouble in getting insights into how spatial data, data mining software and the whole process of data preparation work. Data Mining requires a huge amount of time and effort. The processing of the data and the interpretation of the many different rules produced by the Data Mining applications requires a great deal of time and patience. Data mining rules cannot always be used and is not always the answer to a problem for getting new information out of a large dataset. In the case of the research presented here, the dataset can directly be used by a GIS package to try to find rules by drawing maps and then take it from there to try and find interesting information. Alternatively, statistical analysis may prove to be more productive than data mining. It has been expressed throughout this paper that the quality of the initial census data constrained the extent of knowledge discovery that was possible through this Data Mining effort. It has to be stressed, however, that a large amount of hidden knowledge is contained in the various Data Mining outputs. While traditional query methods require the user to know what questions to ask, Data Mining is successful in pointing out knowledge areas that the researcher might not have thought of to investigate. Nevertheless, the researcher should have a good understanding of the data and a clear purpose with regard to the research. Care should be taken when interpreting the different results,

since the results can often be misleading or distorted, due to the nature of your data.

References

- [1] Adriaan, P., & Zintine, D.(1996) *Data Mining*, Addison Wesley, Harlow:UK.
- [2] MacEachren, A.M. et.al. (Wachowicz, M. & Edsall, R. & Haug, D.) 1999. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *International Journal of Geographical Information Science*, 13 (4):311-334.
- [3] Victor, H.L.& Arndt, H. (2000) *Combining Data Mining and Human Expertise for Making Decisions, sense and policies*. Working Paper, Department of Informatics, University of Pretoria.
- [4] Han, J., & Kamber, M.(2001). *Data Mining, Concepts and Techniques*, Academic Press, U.S.A.
- [5] Koperski, K., 1998. *Spatial Data mining, Spatial Databases*. Retrieved 7 April 2001 from the World Wide Web: <http://www.cs.sfu.ca/people/GradStudents/koperski/personal/research>
- [6] Openshaw, S., et.al. (Turner, A., Turton, I., & Macgill, J.)1999. *Geographical Data Mining*. Retrieved 26 September 2000 from the World Wide Web: <http://www.ccg.leeds.ac.uk/smart/data/gisruk99.htm>.
- [7] Pretotius, J. 2001. *the use of Geographical Information Systems for Knowledge Discovery in Databases*. Working Paper, Department of Informatics, University of Pretoria.
- [8] Schoeman, J.H & Blyth, C, J., 2001. *Data Mining the 1996 South African census data with a view to Knowledge Discovery*. Working Paper, Department of Informatics, University of Pretoria.