

# Application of Spatial Association Rules for Improvement of a Risk Model for Fire and Rescue Services

Věra Karasová, Jukka Matthias Krisp and Kirsi Virrantaus

Helsinki University of Technology, Department of Surveying, Institute of Cartography and Geoinformatics

PO Box 1200, FIN-02015 HUT, Finland

vera.karasova@hut.fi, jukka.krisp@hut.fi, kirsi.virrantaus@hut.fi

WWW home page: <http://www.tkk.fi/Units/Cartography/>

**Abstract.** Throughout this paper, the existing knowledge about Spatial Data Mining (SDM) and more precisely spatial association rule induction is discussed. To test, whether SDM is a useful method for analyzing spatial data stored in an extensive geographical database, this research continues with application of association rule mining on a case study. The core of the data selected for the case study consists of records of incidents, which are located within the Helsinki city center. The goal of the case study is to discover the possible influence of selected geographical objects on the occurrence of incidents.

## 1 Introduction

Due to rapid growth of data collection in recent years, traditional statistical analysis tools are having difficulties with handling the huge volumes of data and the growing complexity of analysis tasks. Moreover, the statistical methods require a broader knowledge of the test data in order to define a principal hypothesis of the analysis. As a consequence, the analyzes become more expensive and time consuming. [Miller and Han, 2001], [Shekhar and Chawla, 2003], [Mannila, 2002]

A new discipline, which concentrates on the manipulation of extensive databases has been introduced as Data Mining (DM). The main goal of DM is to search for deeply hidden information, that can be turned into knowledge for strategic decision making and answering fundamental research questions [Miller and Han, 2001]. Due to the ability of extracting implicit knowledge without any *a priori* stated hypothesis, DM has become a popular tool for data analyzes.

Collected data is not always randomly distributed, independent or stored in relational databases. Considering geographical data, the spatial autocorrelation and complex spatial relations have to be taken into consideration. Although the

DM techniques are very powerful, in order to use them for analyzing spatial data, their potential has to be upgraded to understand core spatial characteristics. Spatial Data Mining (SDM), concerning only analyzes of spatial data, was introduced as a part of the DM discipline. [Shekhar and Chawla, 2003] Even though the requirements of SDM often differs from classical data mining in principal, some SDM techniques tend to adjust classical data mining algorithms instead of designing new ones.

Each SDM technique is developed for analysis of different spatial phenomena. The most often used SDM methods like clustering, trend detection and classification [Ester et al., 2001], [Han et al., 2001], [Shekhar and Chawla, 2003] etc. are derived from spatial statistics, therefore they do not bring any significant innovation to spatial analysis. The only method, that offers a solution for identifying not explicitly stored spatial associations, which is therefore capable of discovering interesting and unexpected relationships, is the association rule mining ([Ester et al., 2001], [Koperski and Han, 1995], [Miller and Han, 2001] etc.).

This paper is based on a literature survey which identifies the core concepts of association rule induction as one of the basic techniques of SDM. This background is a starting point for further theoretical and conceptual analysis. To test the interaction of SDM with real data, the association rule mining is applied to a case study. The main goal of the case study is to present a solution covering the whole process of operations, that are necessary for obtaining desirable results. It must be emphasized that the explanation of this process and description of each of its steps is of main importance to this paper.

In section 2 the core definitions of spatial association rule mining are outlined. The description of the data used in the case study is given in section 3. Section 4 illustrates the three steps of the designed process. The results are evaluated and the study is concluded in section 5.

## 2 Spatial Association Rules

The spatial association rule is a rule denoting certain association relationships among a set of spatial and possibly non-spatial attributes of geographical objects. Those attributes are indicated as predicates which may represent topological relationships between spatial objects, such as *disjoint*, *intersects*, *adjacent.to* etc., they can also hold information about spatial orientation or ordering like *left*, *north*, *east* etc., or specify a distance e.g. *close.to*. [Koperski and Han, 1995]

A spatial association rule is of the form  $X \Rightarrow Y (c \%)$ , where  $X$  is called antecedent and  $Y$  consequent of the rule. The antecedent contains a set of predicates from the exploring database, the consequent only represents one predicate, which is not yet included in the antecedent. The rule itself then reflects an existing relationship between predicates in antecedent and consequent. A measure of

a rule's strength is *confidence* ( $c\%$ ), which indicates that  $c$  percent of the items satisfying the antecedent also satisfies the consequent. A large number of spatial association rules can be derived from an enormous geographical database. However, only a few rules are found interesting. To ensure that only interesting rules are generated, certain additional constraints have to be fulfilled. Those constraints are:

1. *Syntactic Constraints*: These constraints involve restrictions on items that can appear in a rule, e.g. only rules that have a specific item appearing in the consequent are extracted.
2. *Support Constraints*: The *support* of a rule is defined to be the fraction of all considered transactions that satisfy the union of predicates in the consequent and antecedent of the rule. Rule is evaluated as interesting, if its support exceeds the *minsupport* threshold. [Agrawal et al., 1993]

The association rule becomes *strong* when the support is *large*, i.e., no less than the minimum support threshold, and the confidence is *large*, i.e., no less than the minimum confidence threshold. [Agrawal et al., 1993], [Koperski and Han, 1995] To facilitate the analysis, an efficient algorithm is needed to restrict the search space. One of the best known algorithms for mining spatial associations is called *Apriori algorithm* and was developed by Agrawal et al. [Agrawal et al., 1993]. This algorithm works in two steps. In the first step large itemsets (predicate-sets) are determined. These large itemsets contain items which occur frequently together, i.e. the count exceeds the *minsupport* threshold. The second step represents the actual generation of association rules from the large itemsets detected in the first step. [Borgelt and Kruse, 2002]

An important part of the association rule mining and one of the biggest challenges is evaluation of generated rules. Hundreds of rules can easily be obtained from a geographical database. It is obvious that evaluation of all rules one-by-one is impossible for a human expert. To extract only those representing valuable information, some automated techniques are used. The problem is partly solved by extracting only strong rules. However, a weak rule can also be interesting, because it may represent a negation. For example a rule *school*  $\Rightarrow$  *park* with low value of support and confidence indicates, that in an examined area a park is very seldom situated next to a school. Once discovered, this information can help urban planners with designing locations of new parks and playgrounds.

In addition, not all strong rules necessarily hold considerable information. Let's assume a rule *bus\_stop*  $\Rightarrow$  *road* with a confidence over ninety percent. Although this association is very strong, it does not discover any interesting pattern. It is obvious that bus stop is usually placed next to the road.

As denoted, selection of interesting rules is a very complicated process which is strongly dependent on the studied database and desired results. More about problems connected with utilization of support and confidence constraints for rule evaluation is explained in [Borgelt and Kruse, 2002].

The other way of solving problems related to extraction of only important rules is to apply syntactic constraints. By designing a simple template, where the possible appearance of certain items or predicates is stated, only rules fulfilling the constraints are selected from the database independently on the value of the confidence and support. [Borgelt and Kruse, 2002], [Klemettinen et al., 1994]

### 3 Case study

Since the attributes of data stored in various geographical databases differ, the method of extracting association rules needs to be adjusted to the nature of the investigated data. In general, the process of mining geographical data is very extensive. This paper demonstrates one possible solution. The proposed process is explained on a case study situated near the center of Helsinki.

The aim is to determine existing spatial relationships between the location of incidents and specific geographical objects within the study area. The data collected for this study are obtained from two databases. The records about incidents are provided by the Fire and Rescue services in Espoo. The additional objects are extracted from SeutuCD, which is a geographical database of the Finnish Metropolitan area maintained by Helsinki Metropolitan Area Council (YTV). To facilitate the analysis, only eleven object types represented in eleven vector layers are selected from both databases. The objects are bars and restaurants, incidents, kindergartens, main roads, minor roads, motorways, parks and cemeteries, paths, railways, water and waterways. The data selected for the analysis is slightly modified for the security reasons. Each of the objects is characterized by only its unique ID, geographical coordinates and the specification of object type (point, line, polygon). Due to the restriction of attributes, the only subjects of investigation are the spatial relations of those objects. Since various spatial relations can be defined, the only one considered in this research is the immediate neighbourhood of the selected objects.

### 4 Method

The process of generating associations is complex and requires performance of diverse operations. The whole process of identifying significant relations among selected objects is explained in figure 1. The three core steps of the designed process are data pre-processing, transformation to the transaction format and association rule mining. Those steps are symbolized in the figure by red ellipses. The black boxes, connected by pointing arrows represent the particular actions, that have to be performed in each of the three steps. The left side of the schema illustrates the successive changes of the data format together with additional files needed for the analysis. Supporting operations are depicted on the right side of the schema together with the program selected for the extraction of association rules. [Karasová, 2005]

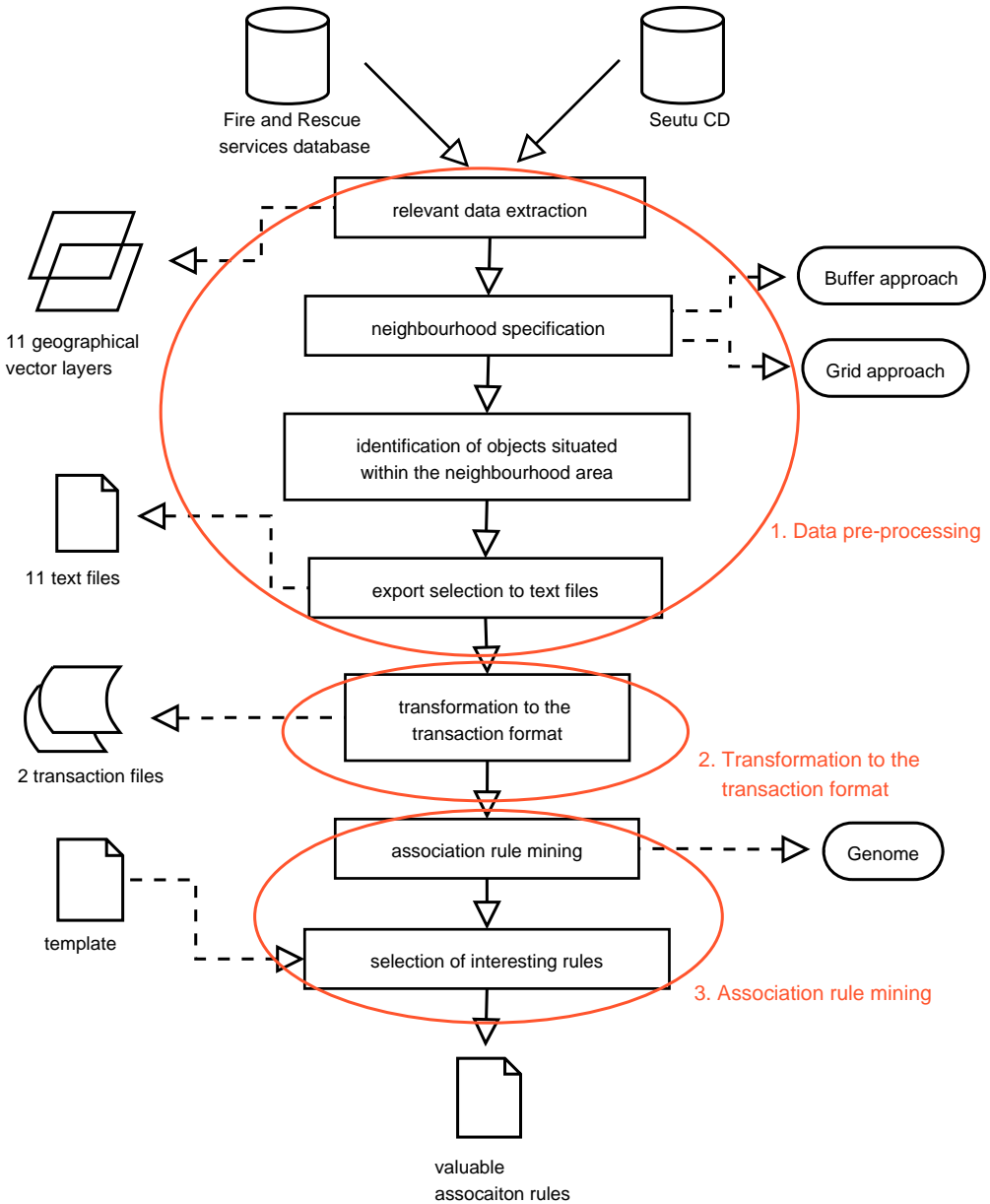
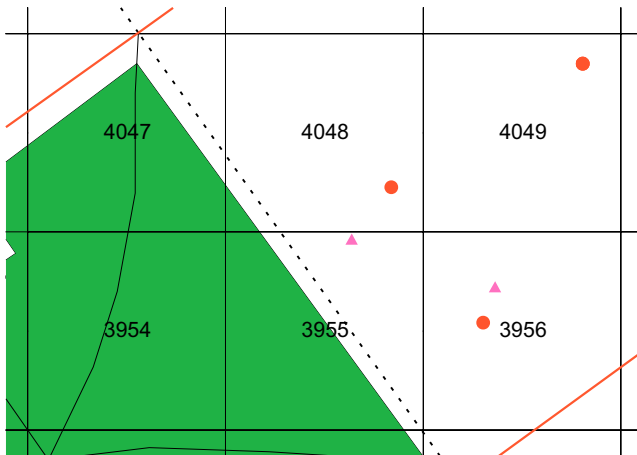


Fig. 1. Schema of the association rule mining process. [Karasová, 2005]

#### 4.1 Data pre-processing

The pre-processing part deals with extraction of relevant data from the two databases and with identification of the proximate neighbourhood among the selected objects. In this research two different approaches are designed for the neighbourhood definition, the grid approach and the buffer approach. The basic idea of the grid approach is based on the vertical-view model introduced by [Estivill-Castro and Lee, 2001]. The original method is modified to fit our extracted data. Each of the eleven layers of the study area are successively overlaid with a rectangular grid with cell size 50m x 50m. Every grid cell identifies a neighborhood and is characterized by a unique ID number. Once an object of a certain vector layer is identified inside of a specific grid cell, the ID number of the cell is depicted and stored in a newly created text file representing the vector layer. For example in figure 2 the text file representing incidents consists of cells 3956, 4048, 4049. By the same process the text file identifying parks and cemeteries contains the ID's of cells 3954, 3955, 4047. The files representing remaining objects are filled in the same way. The output of this operation is eleven text files containing only ID numbers of relevant cells.



**Fig. 2.** Numbered grid cells define neighbourhood areas. Red circle represents incident locations. Bars and restaurants are illustrated by the ginger pink triangles. Green polygons depict parks and cemeteries, red lines main roads black dashed lines minor roads, finally thin black line represents walking paths.

A more accurate way of identifying neighbourhood areas is the buffer approach which generates circular buffers with a radius of 50m. Since our main interest is concentrated on the incidents, the buffers are created only around them. The identification of neighbourhood objects to incidents is similar to the grid approach. From figure 3 the only related object to incident 54 is a main road, the incident with ID number 512 neighbours with a main and a minor road and a

kindergarten. The results of the buffer approach are stored in ten text files, the incident file is excluded. Each file, similar to the grid approach, represents a certain object layer and contains ID numbers of incidents, to which the particular object is identified as related.



**Fig. 3.** A circular yellow buffer is created around numbered incidents represented by a red circle. Bars and restaurants are illustrated by the ginger pink triangles. Blue polygons depict water, red lines main roads, black dashed lines minor roads and the last line denotes railway. Finally kindergartens are represented by a gold square.

## 4.2 Transformation to the transaction format

The second step of the process integrates all the separate files into one representing itemsets. The basic idea of the integration is the same in both approaches, however, one more operation is added when transforming data obtained by the buffer approach.

The integration process for files obtained with the grid approach is depicted in figure 4. The top part of the figure represents three generated text files. Each input file is assigned a number according to the order to the integration algorithm. For instance, the file representing railways is given number 1, because it is detected first. Numbers in the columns denote the cell ID's. The steps of the integration algorithm are:

1. Check every cell ID number of the grid.
2. If the ID number occurs in a file, classify the cell according to the file of origin, in our example (1, 2 or 3).
3. Add the classified cell as an item to the Results file.
4. If the same ID number exists in a different file, add the file number to the already created itemset in the Results file.
5. Save the Results file.

For integration of buffer data, the algorithm is extended with one more step. After the *Results* file is filled, one more item is added to every itemset. All the

RAILWAY(1)	INCIDENTS(2)	BARS(3)
2	10	1
45	45	8
48	46	16
50	50	23
132	133	50
145		133
159		165
		181

RESULTS
3
1
3
2
3
3
1 2
1
2
1 2 3
1
2 3
1
1
3
3

**Fig. 4.** Integration algorithm for the data obtained by the grid approach. An example is highlighted in red. After several passages through the files, cell number 50 is detected in the railway file. The cell is classified as number 1, because 1 is the label of a railway file. Consequently, a new itemset is created in the *Results* file. The same number (50) is found in file number 2, i.e. incidents. Therefore, the algorithm adds the item to the already existing itemset. Now the itemset contains two items 1 and 2, railway and incident. Finally, the same cell number is detected also in the third file representing bars and restaurants. The cell is again classified by the number of the file and added to the itemset. The final itemset is of the form 1, 2, 3 and states: *In one location within the study area, railway, incident and a bar or restaurant are identified as adjacent objects.* [Karasová, 2005]

relations in the buffer approach are specified to incidents only. However, the *Results* file does not, unit now, contain any information about it. Therefore, the additional item, substituting incidents, respectively the buffers around incidents, makes the itemsets complete.

### 4.3 Association rule mining

Once the transaction file is obtained, the application of association rule mining is straightforward. The aim of this research is not to implement the *Apriori* algorithm, we concentrate on possible application of an already existing tool. The program used for generation of association rules is designed by Borgelt and its implementation is explained in [Borgelt and Kruse, 2002]. A graphical user interface was developed by Togaware [Togaware, 2005] as part of the *Gnome Data Mine* tool, and can be downloaded from [Gnome, 2005].

The previous two steps describe the conversion of provided data to the format, which is understandable for the selected software. We are aware that large amounts of rules can be generated from the two transaction files. To obtain only valuable rules, the extraction has to be restricted. Therefore, three constraints are defined:

- Minsupport
- Minconfidence
- Syntactic constraint

Because rules with low confidence can hold relevant information, the *minconfidence* threshold is set to zero. With respect to the *minconfidence*, the *minsupport* is also equal to zero. With these settings, all existing rules are extracted from the itemsets. However, not all of them are interesting. The other way of solving problems related to extraction of only important rules is to apply the syntactic constraints by designing a simple template, where the possible appearance of certain items is stated. Only rules fulfilling the constraint are selected from the database independently on the value of the confidence and support. In this case the designed template limits the number of rules to only those, containing incidents. The three constraints are applied to both transaction files with the same values.

## 5 Results and Conclusions

After all pre-defined constraints are set in the Gnome data mine tool, the association rule mining is applied separately on both transaction files. The following rules extracted from the transaction file obtained by the grid approach are identified:

*bars and restaurants*  $\Rightarrow$  *incidents* (1.7%; 40.0%) (1)

*incidents*  $\Rightarrow$  *main roads* (2.2%; 30.4%) (2)

*incidents*  $\Rightarrow$  *minor roads* (1.7%; 24.1%) (3)

*motorway*  $\Rightarrow$  *incidents* (0%; 2.9%) (4)

*incidents*  $\Rightarrow$  *water* (0.4%; 5.7%) (5)

The first number between brackets represents the support and the second the confidence of the rule. Rule 1 is the strongest from all the generated rules and states: *An incident has occurred in a neighbourhood of 40 % of all bars and restaurants within the Helsinki city center during the studied time period.* This means that there is a high probability that the presence of bars and restaurants strongly affects the occurrence of an incident. Rules 2 and 3 show an association between incidents and two specific road classes. Rules 4 and 5 are examples of detected negation. Rule 4 shows that accidents on a motorway are not very common. However, this rule does not have an important meaning for this particular area, because the motorway passes only through a negligible part of the Helsinki center. A more illustrative example of expressing negation between incidents and spatial object is rule 5. Although the water covers more than 50 % of the study area, the association between water and incidents is not prominent. Therefore the water does not have a strong impact on incidents.

The results obtained with the buffer transaction file can not be, in this stage, used as an adequate source of information. Since the buffer is created only around incidents, all the itemsets existing in the Results file are only incident related. By mining this file the existing relationships between incidents and other objects are correctly detected, however, the measures of the extracted rules can not be compared to the results obtained by the grid approach. Nevertheless, the buffer and grid approaches are independent on each other, therefore the rules obtained from the buffer transaction file can be taken as supporting results. Since significant rules generated from the buffer transaction file are the same to those extracted from the grid transaction file, the information about discovered relations is verified.

To illustrate the possible use of SDM techniques as an alternative approach to commonly used spatial analysis methods, we demonstrated spatial association rule mining on a case study. A new application was not implemented, a tool, originally designed for classical data mining was used for the analysis. We are aware, that the proposed process does not offer a general solution, which can be applicable to any geographical database. Moreover, the identification of the object neighbourhood is rather simple and the amount of data used in the case study is heavily restricted. However, the process is open to further improvements. Some of them are proposed in [Karasová, 2005]. The aim of this research

is to introduce the core concepts of SDM. By the case study, we accomplished that application of SDM is effective for detecting strong relationships among geographical objects.

## References

- [Agrawal et al., 1993] Agrawal R., Imielinski T. and Swami A., Mining association rules between sets of items in large databases, proceedings of ACM-SIGMOD International Conference Management of Data, pages 207-016, 1993
- [Borgelt and Kruse, 2002] Borgelt Ch. and Kruse R., Induction of Association Rules: Apriori Implementation, proceedings of 14th Conference on Computational Statistics, 2002
- [Ester et al., 1997] Ester M., Kriegel H.-P. and Sander J., Spatial Data Mining: A Database Approach, proceedings of 5th International Symposium on Advances in Spatial Databases, pages 47-66, 1997
- [Ester et al., 2001] Ester M., Kriegel H.-P. and Sander J., Algorithms and applications for spatial data mining, in Geographic data mining and knowledge discovery, Miller H. J. and Han J., Taylor & Francis, ISBN 0-415-23369-0, 2001
- [Estivill-Castro and Lee, 2001] Estivill-Castro V. and Lee I., Data Mining Techniques for Autonomous Exploration of Large Volumes of Geo-referenced Crime Data, proceedings 6th International Conference on Geocomputaion, GeoComputation CD-ROM, ISBN 1864995637, 2001
- [Gnome, 2005] URL: <http://www.togaware.com/datamining/gdatamine/gdmapriori.html> (accessed 22.3.2005)
- [Han et al., 2001] Han J., Kamber M. and Tung A. K. H., Spatial clustering methods in data mining, in Geographic data mining and knowledge discovery, Miller H. J. and Han J., Taylor & Francis, ISBN 0-415-23369-0, 2001
- [Karasová, 2005] Karasová V., Spatial data mining as a tool for improving geographical models, Master's thesis, Helsinki University of Technology, Department of Surveying, 2005 URL: <http://www.tkk.fi/Units/Cartography/theses/master/index.html>
- [Koperski and Han, 1995] Koperski K. and Han J., Discovery of Spatial Association Rules in Geographic Information Databases, proceedings of 4th International Symposium on Large Spatial Databases, pages 47-66, 1995
- [Klemettinen et al., 1994] Klemettinen M., Mannila H., Ronkainen P., Toivonen H. and Inkeri Verkamo A., Finding Interesting Rules from Large Sets of Discovered Association Rules, proceedings of 3rd International Conference on Inforamtion and Knowledge Management, pages 401-408, 1994
- [Mannila, 2002] Mannila H., Local and Global Methods in Data Mining: Basic Techniques and Open Problems, proceedings of 29th International Colloquium on Automata, Languages and Programming, Lecture Notes on Computer Science, pages 57-68, Springer-Verlag, 2002
- [Miller and Han, 2001] Miller H. J. and Han J., Geographic data mining and knowledge discovery, An overview, in Geographic data mining and knowledge discovery, Miller H. J. and Han J., Taylor & Francis, ISBN 0-415-23369-0, 2001
- [Shekhar and Chawla, 2003] Shekhar S. and Chawla S., Introduction to Spatial Data Mining, in Spatial Databases: A tour, Prentice Hall, ISBN 013-017480-7, 2003
- [Togaware, 2005] Togaware, URL: <http://www.togaware.com/index.html> (accessed 22.3.2005)