

# Data Mining Methods for the Analysis of Large Geographic Databases

Krzysztof Koperski and Jiawei Han

*Spatial data mining, i.e., discovery of interesting, implicit knowledge in spatial databases, is an important task for understanding and use of spatial data- and knowledge-bases. Statistical analysis has been the main method used for analyzing spatial data. Unfortunately, it has a number of weaknesses. In this paper, a number of methods based on knowledge discovery techniques for large databases are presented. This methods may overcome some of the weaknesses of statistical analysis. Our study is focused on efficient method for mining strong spatial association rules in geographic information databases. A spatial association rule is a rule indicating certain association relationship among a set of spatial and possibly some non-spatial predicates. For example, a rule "80% of gas stations in rural areas are close to highways" is a spatial association rule. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database and strong implication relationships.*

## INTRODUCTION

With wide applications of satellite and remote sensing technologies and automatic data collection tools, tremendous amounts of spatial and non-spatial data have been collected and stored in **large spatial databases**. The extraction and comprehension of the knowledge implied by the huge amount of spatial data, though highly desirable, pose great challenges to currently available GIS technologies. This situation demands new technologies for **knowledge discovery in large spatial databases**, or **spatial data mining**, that is, *extraction of implicit knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial databases*.

Statistical spatial analysis is widely used technique for analyzing spatial data. Statistical analysis works well with numerical data and enables optimization and building models. Unfortunately, statistical analysis has some limitations including poor dealing with symbolic data like names, high computational complexity and others.

In this paper we present some methods which extend existing data mining techniques toward spatial data mining. We also propose and study an efficient method for mining strong spatial association rules in large geographic

databases. A strong rule indicates that the patterns in the rule have relatively frequent occurrences in the database (i.e., **large support**) and strong implication relationships (i.e., **high confidence**). Data mining process is initialized by a user who defines in a query types of relations and classes of objects for description. Our algorithm enables efficient discovery of interesting spatial association rules in large geographic databases, which may have high application potential in practice.

## METHODS

### Statistical Analysis

Currently statistical spatial analysis is the most common technique for analyzing spatial data [4]. Statistical methods handle well numerical data, contain a large number of algorithms, have a strong possibility of getting models of spatial phenomena, and allow optimizations. Statistical analysis sometimes requires the assumptions regarding to statistical independence of spatially distributed data. Such assumptions are often unrealistic due to the influence of neighboring regions. To deal with such problems, spatial models can include trend surface or dummy variables. If data in one region are influenced by features of neighboring regions, the analyst may fit a regression model with a spatial lagged forms of the dependent variables. Statistical analysis also deals poorly with symbolic data like names. Statistical approach requires a lot of domain and statistical knowledge. Thus, it should be performed by domain experts with the experience in statistics. Statistical methods also does not work well with incomplete or inconclusive data. Another problem related to statistical spatial analysis is expensive computation of the results. To overcome some of the weaknesses of statistical analysis new methods have been proposed for analyzing data which is stored in relational databases [3].

### Generalization-based Methods

In this methods, knowledge is summarized in a form of relationships between spatial and non-spatial attributes at generalized high concept level. In [7], attribute-oriented

### Non-spatial dominant generalization

extract region  
from precipitation-map  
where province = "B.C." and  
and period = "spring" and year = 1990  
in relevance to precipitation and region

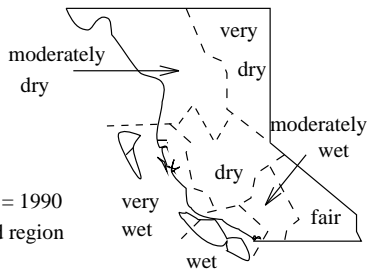


Figure 1: An example of the query using non-spatial dominant generalization method.

induction technique was used for knowledge discovery in spatial databases. Two kinds of hierarchies were constructed. One is the concept hierarchy describing non-spatial data; the second hierarchy describes spatial data. Non-spatial hierarchy provides the concept of generalizing data to high level concepts. For example, numerical data can be generalized to ranges or descriptive high level concepts (e.g.,  $-9^{\circ}\text{C}$  to a range value " $-10\text{ to }0^{\circ}\text{C}$ " or *cold*), and symbolic values to higher level concepts (e.g., *potatoes* and *beets* to *vegetables*). By doing so, low level distinctive values may be generalized to identical high level values, and such high-level identical values among different tuples can be merged together with their spatial pointers clustered into one slot in the spatial attribute. Spatial hierarchies can be defined by administrative division or spatial storage structures like quad-trees.

Two algorithms were proposed: *non-spatial data dominated generalization*, and *spatial dominated generalization*. In the first step of both algorithms, data related to the user specified query are collected. Then, in *non-spatial data dominated algorithm*, attribute-oriented induction is performed on non-spatial data by (a) climbing concept hierarchy, (b) attribute removal, and (c) merging identical tuples. Induction is continued until every attribute is generalized to the desired level. In the last step, neighboring areas with the same generalized attributes are merged. The result of the query can be presented as a map with small number of regions with high level descriptions: (Figure 1-1).

For *spatial data dominated generalization*, the task-relevant spatial data is generalized first. It is done by clustering spatial objects until a desired concept level is reached. For each generalized region, non-spatial attributes are generalized in all regions until proper description for all regions can be done. Finally, the output of the query presents the description of spatial regions by a small number of predicates: (Figure 1-2).

### Spatial dominant generalization

extract characteristic rule  
from temperature-map  
where province = "B.C." and  
period = "summer" and year = 1990  
in relevance to region and temperature.

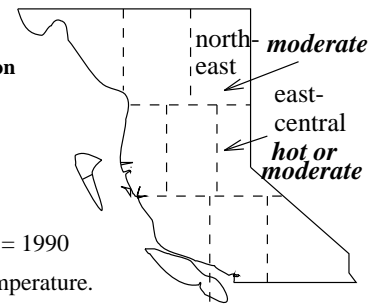


Figure 2: An example of the query using spatial dominant generalization method.

## Mining using clustering techniques

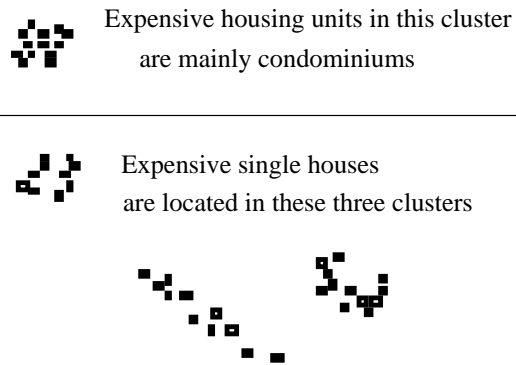


Figure 3: Two examples of queries using spatial and non-spatial dominant generalization methods.

## Cluster Analysis

Combination of attribute-oriented induction with clustering methods provides a possibility of describing spatial behavior of similar objects or to determine characteristic features of distinct clusters.

In [8], two techniques were derived: *spatial-dominant* and *non-spatial-dominant* algorithms (Figure 1-3). The *spatial-dominant* method classifies all task-relevant spatial objects (such as points) into clusters using an efficient clustering algorithm based on sampling. Then it performs an attribute-oriented induction on non-spatial description of objects in each cluster to extract rules describing general properties of a cluster. The *non-spatial-dominant* method first generalizes non-spatial attributes of query-related objects to high concept levels. For example, expensive housing units can be generalized to single houses, mansions and condominiums. Finally, the program clusters the spatial objects with the same non-spatial descriptions.

## Spatial Association Rules

Agrawal, et. al. [1] proposed association rules for discovery of interdependencies in transactions databases. If this method is used for knowledge discovery in supermarket database a user can discover rules in form of “ $W \rightarrow B$  ( $c\%$ )”, where  $W$  and  $B$  are sets of attribute values. This rule is explained as, “if a pattern  $W$  appears in a transaction, there is  $c\%$  possibility (confidence) that the pattern  $B$  holds in the same transaction”. For example, the rule “ $bread \rightarrow butter$  (90%)”, which states that 90% of customers who buy bread also buy butter, is an association rule. Our study extends this method towards discovery of spatial relations.

### Definitions

A **spatial association rule** is a rule in the form of

$$P_1 \wedge \dots \wedge P_m \rightarrow Q_1 \wedge \dots \wedge Q_n. \quad (c\%) \quad (1)$$

where at least one of the predicates  $P_1, \dots, P_m, Q_1, \dots, Q_n$  is a spatial predicate, and  $c\%$  is the *confidence* of the rule which indicates that  $c\%$  of objects satisfying the antecedent of the rule will also satisfy the consequent of the rule [6].

There are various kinds of spatial predicates that could constitute a spatial association rule. Some examples are: topological relations like *intersects*, *overlap*, *disjoint*, etc.; spatial orientations like *left\_of*, *west\_of*, etc, distance information, such as *close\_to*, *far\_away*, etc.

The **support** of a conjunction of predicates,  $P = P_1 \wedge \dots \wedge P_k$ , in a set  $S$ , denoted as  $\sigma(P/S)$ , is the number of objects in  $S$  which satisfy  $P$  versus the cardinality (i.e., the total number of objects) of  $S$ . The **confidence** of a rule  $P \rightarrow Q$  in  $S$ ,  $\varphi(P \rightarrow Q/S)$ , is the ratio of  $\sigma(P \wedge Q/S)$  versus  $\sigma(P/S)$ , i.e., the possibility that  $Q$  is satisfied by a member of  $S$  when  $P$  is satisfied by the same member of  $S$ .

Since most people are interested in rules with large supports and high confidence, two kinds of thresholds: *minimum support* and *minimum confidence*, can be introduced. Moreover, since many predicates and concepts may have strong association relationships at a relatively high concept level, the thresholds should be defined at different concept levels. For example, it is difficult to find regular association patterns between a *particular house* and a *particular beach*, however, there may be strong associations between many *expensive houses* and *luxurious beaches*. Thus, it is expected that many spatial association rules are expressed at a relatively high concept level.

A set of predicates  $P$  is **large** in set  $S$  at level  $k$  if the support of  $P$  is no less than its minimum support threshold  $\sigma_k^l$  for level  $k$ , and all ancestors of  $P$  from the concept hierarchy are large at their corresponding levels.

The confidence of a rule “ $P \rightarrow Q/S$ ” is **high** at level  $k$  if its confidence is no less than its corresponding minimum confidence threshold  $\varphi_k^l$ .

A rule “ $P \rightarrow Q/S$ ” is **strong** if the predicate “ $P \wedge Q$ ” is *large* in set  $S$  and the *confidence* of “ $P \rightarrow Q/S$ ” is high.

### Algorithm

The general distributed algorithm for the mining process can be summarized in the following.

**Input:** consists of a spatial database, a mining query, and a set of thresholds.

1. The database consists of three parts: (1) a spatial database, *SDB*, containing a set of spatial objects; (2) a relational database, *RDB*, describing non-spatial properties of spatial objects; and (3) a set of concept hierarchies.
2. The query consists of: (1) a class  $S$  of objects being described, (2) a set of task-relevant subclasses for spatial objects  $C_1, \dots, C_n$ , and (3) set of relevant relations/predicates.
3. Three thresholds: (1) the minimum support, (2) the minimum confidence for each level of concept hierarchies; and (3) the distance threshold specifying the maximal distance between the objects to satisfy *close\_to* predicate.

**Output:** Strong multiple-level spatial association rules for the relevant sets of objects and relations.

**Method:** The procedure for mining spatial association rules is the following:

- 1: *map1* := map of the class being described
- 2: For all subclasses  $C_i$  of relevant objects do concurrently:
- 3:     *map2* := map of a subclass of relevant objects
- 4:     *Predicates\_for\_Ci* := Intersection(*map1*, *map2*, *dist*);
- 5:     *Predicates\_DB*  $\cup$  = *Predicates\_for\_Ci*;
- 6: Find\_large\_predicates\_and\_mine\_rules(*Predicates\_DB*);

Step 4 of the algorithm is accomplished by execution of efficient multilevel algorithm for computation of spatial joins [2]. Predicates are stored in an extended relational database *Predicates\_DB*, which allows an attribute value to be either a single value or a set of values. Every row of the *Predicates\_DB* is a description of a single object  $S_i$  from the the class  $S$  of objects being described. Description consists of objects  $C_{jk}$  and relations which satisfy task relevant predicates. For example, row related to Stanley Park may include restaurant, zoo, main road, inlet, lake etc. Step 6 is executed by algorithm for discovery of multilevel association rules in transaction databases

[5]. This algorithm finds large predicates by counting the number of occurrences of predicates in the database and based on this information derives strong rules. For example, if predicate *close\_to(x, A)* occurs in 100 rows, predicate *close\_to(x, B)* occurs in 90 rows, and both predicates *close\_to(x, A)* and *close\_to(x, B)* occur together in 80 rows, then the rule “*close\_to(x, A) → close\_to(x, B)* (80%)” may be derived. After finding large predicates on high levels of concept hierarchies the algorithm tries to find large predicates and rules on lower levels. For example, highways may be specialized into interstate highways and state highways.

### Implementation

The algorithm, was implemented using C and SR languages. The program was tested on two databases consisting of the description of Washington state and Hawaii from Census Bureau’s TIGER/Line(TM) files. It was executed on the farm of SPARCstations 5. The execution time for the average query is about 80s for a database of size about 22MB and about 12s for database of size 4.5MB. On the top level of the concept hierarchies the algorithm may discover that 37% of airports and airfields in Washington state are close to railroads. On the lower level of concept hierarchy it finds that 75% of airports and airfields which are close to U.S. state highways are also close to railroad main tracks.

### CONCLUSION

We investigated and implemented a number of algorithms extending spatial analysis by using methods for knowledge discovery in large databases. Our study shows an effective method for user guided data mining which can be used for analyzing of large spatial databases. Discovery of spatial association rules may disclose interesting relationships among spatial and/or non-spatial data in large spatial databases and thus, it represents a new and promising direction in spatial data mining.

### Acknowledgments

This research was supported in part by the research grant NSERC-A3723 from the Natural Sciences and Engineering Research Council of Canada and the research grant NCE/IRIS-IC2 from the Networks of Centers of Excellence of Canada.

### References

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases.

In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, pp. 207-216, Washington, D.C., May 1993.

[2] T. Brinkhoff, H. P. Kriegel, R. Schneider, B. Seege. Multistep Processing of Spatial Joins. In *Proc. 1994 ACM-SIGMOD Conf. Management of Data*, Minneapolis, Minnesota, pp. 197-208, May 1994.

[3] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. *Advances in Knowledge Discovery and Data Mining*, AAAI/MIT Press, 1995.

[4] S. Fotheringham, and P. Rogerson. *Spatial Analysis and GIS*, Taylor and Francis, 1994.

[5] J. Han, and Y. Fu. Discovery of Multiple-Level Association Rules from Large Databases in *Proc. 1995 VLDB*, Zurich, Switzerland, pp. 420-431, Sept. 1995.

[6] K. Koperski, and J. Han. Discovery of Spatial Association Rules in Geographic Information Databases. In *Advances in Spatial Databases (Proc. 4th Symp. SSD'95)*, pp. 47-66, Portland, ME, August 1995.

[7] W. Lu, J. Han, and B. C. Ooi. Discovery of General Knowledge in Large Spatial Databases. In *Proc. Far East Workshop on Geographic Information Systems* pp. 275-289, Singapore, June 1993.

[8] R. Ng, and J. Han. Efficient and effective clustering method for spatial data mining. In *Proc. 1994 Int. Conf. VLDB*, pp. 124-155, Santiago, Chile, Sept. 1994

□

*Krzysztof Koperski* is a Ph.D. candidate in the School of Computing Science at the Simon Fraser University at Burnaby B.C. His major research focus is on spatial data mining, spatial reasoning and knowledge discovery in multimedia databases.

*Jiawei Han* is a professor of Computing Science in Simon Fraser University. His major research interests include database and knowledge-base systems, knowledge discovery in databases, deductive and object-oriented databases, spatial and multi-media databases, logic programming, and artificial intelligence. He has served or is currently serving in the program committees of over 20 international conferences, including ACM-SIGMOD'96, VLDB'96 and KDD'96 (Program Committee co-chairman).

School of Computing Science  
Simon Fraser University  
Burnaby, B.C., Canada V5A 1S6  
e-mail: {koperski, han}@cs.sfu.ca  
phone: (604) 291-4411  
fax: (604) 291-3045