A Framework for Regional Association Rule Mining in Spatial Datasets

Wei Ding,* Christoph F. Eick, Jing Wang
Computer Science Department
University of Houston
{wding, ceick, jwang29}@uh.edu

Xiaojing Yuan
Engineering Technology Department
University of Houston
xyuan@uh.edu

Abstract

The immense explosion of geographically referenced data calls for efficient discovery of spatial knowledge. One of the special challenges for spatial data mining is that information is usually not uniformly distributed in spatial datasets. Consequently, the discovery of regional knowledge is of fundamental importance for spatial data mining. This paper centers on discovering regional association rules in spatial datasets. In particular, we introduce a novel framework to mine regional association rules relying on a given class structure. A reward-based regional discovery methodology is introduced, and a divisive, grid-based supervised clustering algorithm is presented that identifies interesting subregions in spatial datasets. Then, an integrated approach is discussed to systematically mine regional rules. The proposed framework is evaluated in a real-world case study that identifies spatial risk patterns from arsenic in the Texas water supply.

1. Introduction

The immense explosion of geographically referenced data calls for efficient discovery of spatial knowledge. The goal of spatial data mining is to automate the extraction of interesting, useful but implicit spatial patterns [10, 16, 18, 6, 1]. One of the special challenges for spatial data mining is that information is usually not uniformly distributed in spatial datasets. It has been pointed out in literature [8, 12, 15] that "*whole map statistics are seldom useful*", that "*most relationships in spatial data sets are geographically regional, rather than global*" and that, "*there is no average place on the Earth's surface*" – a county is not a representative of a state, and a state is not a representative of a country. Therefore, it is not surprising that domain experts are most interested in discovering hidden patterns at a regional scale rather than a global scale [8, 12]. Consequently, the discovery of regional knowledge is of fundamental importance for spatial data mining.

However, most of the current data mining techniques are ill-prepared for discovering regional knowledge. Regional patterns frequently fail to be discovered due to insufficient global confidence and/or support. Furthermore, for a given dataset there is a non-finite number of subregions. This raises the questions on how to measure the interestingness of a set of regions and how to identify regions using a given measure of interestingness.

In this paper, we propose a novel framework to mine regional association rules based on a given class structure. A reward-based regional discovery methodology is introduced, and a new divisive, grid-based supervised clustering algorithm is presented that identifies interesting subregions in spatial datasets. Then, an integrated approach is presented to systematically mine regional rules. The proposed framework is evaluated in a real-world case study that identifies spatial risk patterns from arsenic in Texas water supply. This paper is organized as follows. Section 2 introduces our region discovery framework and Section 3 describes region discovery algorithm and association rule mining algorithm. Section 4 presents the results of the case study and Section 5 concludes the paper.

2. An Integrated Framework for Regional Association Rule Mining

There are two phases in the proposed integrated framework for regional association rule mining:

1. Phase I: Discover and identify interesting subregions. A supervised clustering algorithm using multi-resolution grids divides the whole dataset into a number of non-overlapping spatial subregions. In this phase, there are two challenges: how to measure the interestingness of a set of regions; then given a measure of interestingness, how to identify subregions.

2. Phase II: Spatial association rule mining for each identified subregion. The subregions are considered one at a time and all frequent itemsets for that region are generated. Regional association rules are then constructed

[*] Also, Computer Science Department, UH-Clear Lake.

from these frequent itemsets. The resulting rules are examined. In the case that the results are unsatisfactory for a particular region this feedback will be used to fine tune parameters of the regional discovery algorithm and association rule mining algorithm.

2.1. Problem Formulation

Let $\mathbb{D}$ be a spatial dataset, and $S = \{s_1, s_2, ..., s_l\}$ be a set of spatial attributes, $A = \{a_1, a_2, ..., a_m\}$ be a set of non-spatial attributes, and $CL = \{cl_1, cl_2, ..., cl_n\}$ be a set of class labels. Let

$$\begin{aligned} I &= S \cup A \cup CL \\ &= \{s_1, s_2, ..., s_l, a_1, a_2, ..., a_m, cl_1, cl_2, ..., cl_n\} \end{aligned}$$

be the set of all items in $\mathbb{D}$. Continuous attributes are transformed into nominal attributes. Let $T = \{t_1, t_2, ..., t_N\}$ be the set of all the transactions. T can be represented as a relational table, which contains N tuples conforming to the schema I (I contains $l+m+n$ number of items). Thus an item $i \in I$ is a binary variable whose value is 1 if the item is present in t_i $(i = 1, ..., N)$ and 0 otherwise. Consequently, the set of transactions T is classified based on the given class structure CL.

Our framework employs a class-guided generation of association rules that sheds more light on the patterns related to the given class structure. We define such rules as *supervised association rules*. The formal definition is:

Definition 1 A **supervised association rule** r is of the form $P \rightarrow Q$, where $P \subseteq I$, $Q \subseteq I$, *and* $(P \cup Q) \cap CL \neq \emptyset$.

The rule r holds in the $\mathbb{D}$ with support *sup* and confidence *con* where

$$\begin{aligned} sup(P \rightarrow Q) &= \frac{\sigma(P \cup Q)}{N}, \\ con(P \rightarrow Q) &= \frac{\sigma(P \cup Q)}{\sigma(P)}. \end{aligned}$$

The support count is defined as $\sigma(\alpha) = |\{t_i | \alpha \subseteq t_i,\ t_i \in T\}|$, $(i = 1, ..., N)$, where $|\ .\ |$ denotes the number of elements in a set. A supervised association rule is *strong* if it satisfies user-specified minimum support (*min_support*) and minimum confidence (*min_confidence*) thresholds.

Given these definition, the problem of regional association rule mining can be defined as:

Find: interesting regions and supervised association rules from each discovered region.

Given: a set of items I, a classified transaction set T, a fitness function for the measure of interestingness (see section 2.2), minimum cell size threshold *min_cell_size* for region discovering algorithm (see section 3.1), minimum support threshold *min_support* and confidence threshold *min_confidence*.

2.2. Measuring the Interestingness of a Set of Regions

We define a region as a surface that contains a set of spatial objects. $EXT(R)$, the extension of R, denotes the objects belonging to a region R. A region should be contiguous, that is, for each pair of objects belonging to the same region, there always must be a path within this region that connects them. Consider a global region R, a dataset $\mathbb{D}$, where $\mathbb{D} = EXT(R)$, and an underlying class structure CL, our region discovery algorithm employs a reward-based evaluation scheme that evaluates the quality of the generated subregions. The fitness function, which evaluates the quality of the generated subregions $R_X = \{R_1, ..., R_m\}$, is defined as the sum of the rewards obtained from each subregion R_i $(i = 1..m)$ (Equation 1).

$$\begin{aligned} q(R_X) &= \sum_{i=1}^{m} reward(R_i) \qquad (1) \\ &= \sum_{i=1}^{m} (interestingness(R_i) \times |R_i|^{\beta}),\ where\ \beta > 1. \end{aligned}$$

We find subregions $R_1, ..., R_m$ such that:

1. The subregions are disjoint: $EXT(R_i) \cap EXT(R_j) = \emptyset$, $i \neq j$.

2. $R_X = \{R_1, ..., R_m\}$ maximizes $q(R_X)$.

3. The generated subregions are not required to be exhaustive with respect to R, that is, $EXT(R_1) \cup ... \cup EXT(R_m) \subseteq EXT(R)$.

4. $R_1, ..., R_m$ are ranked based on the reward each region receives. Subregions that receive low rewards or non-rewards are frequently discarded.

This evaluation scheme encourages combining small regions into larger ones if the rewards of the combined regions do not decrease. Consequently, $q(R_X)$ uses $|R_i|^{\beta}$, the region size $|R_i|$ with parameter $\beta > 1$, to increase the value of the fitness nonlinearly and favor a region with more objects.

In this paper, we adopt a single measure of interestingness to find *hotspots* and *coldspots* that were developed and proved to be effective in our previous work [5]. The measure is based on a class of interest $cl \in CL$. It rewards regions in which the density of class cl deviates from its prior

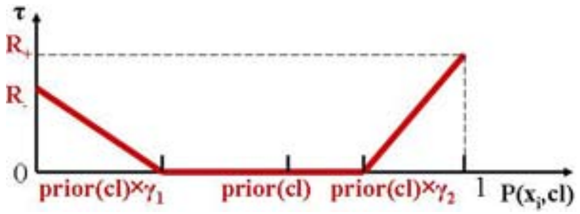

Figure 1. Measure of interestingness τ when $\eta = 1$

probability: A region is a *hotspot* (or *coldspot*) if its density with respect to class cl is significantly higher (or lower) than the expected probability.

Let N denotes number of objects in a dataset $\mathbb{D}$, x_i the i_{th} cluster, and $X = \{x_1, x_2, ..., x_k\}$ a clustering solution consisting of clusters x_1 to x_k. Each cluster corresponds to a subregion $x_i = EXT(R_i)$, $i = 1..k$. The fitness function $q(X)$ (Equation 2) is defined as

$$q(X) = \sum_{i=1}^{k} \tau(P(x_i, cl), prior(cl), \gamma_1, \gamma_2, R_+, R\_, \eta) \times (\frac{|x_i|}{N})^{\beta} \quad (2)$$

The function of interestingness τ (Equation 3) is calculated based on $P(x_i, cl)$ and $prior(cl)$, with the following parameters: η, γ_1, γ_2, R_+, $R\_$, where $\eta > 0$, $\gamma_1 \leq 1 \leq \gamma_2$, $0 \leq R_+$, $R_- \leq 1$. $P(x_i, cl)$ is the probability of objects in cluster x_i belonging to the class of interest cl, and $prior(cl)$ is the probability of objects in datasets $\mathbb{D}$ with respect to the class cl. R_+ and $R_-$are the maximum rewards for hotspot and coldspot respectively.

$$\tau(P(x_i, cl), prior(cl), \gamma_1, \gamma_2, R_+, R_-, \eta) = \quad (3)$$
$$\begin{cases} \left[\frac{prior(cl)\times\gamma_1 - P(x_i,cl)}{prior(cl)\times\gamma_1} \times R_-\right]^{\eta} & if\, P(x_i, cl) < priori(cl) \times \gamma_1 \\ \left[\frac{P(x_i,cl) - prior(cl)\times\gamma_2}{1 - prior(cl)\times\gamma_2} \times R_+\right]^{\eta} & if\, P(x_i, cl) > priori(cl) \times \gamma_2 \\ 0 & otherwise \end{cases}$$

The parameter η determines how quickly the reward grows to the maximum reward (either R_+ or R_-). If η is set to 1, the reward function changes linearly, as shown in Figure 1. In general, the larger value for η, the higher rewards for purer clusters. $prior(cl) \times \gamma_1$ and $prior(cl) \times \gamma_2$ determines the thresholds based on which a reward is given to a subregion.

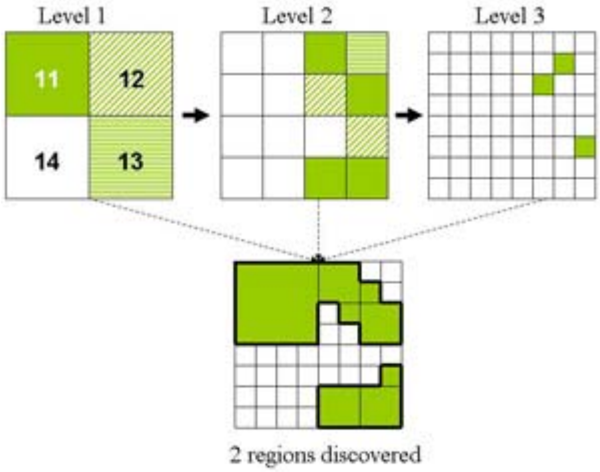

Figure 2. A sample example of running the SCMRG algorithm.

3. Algorithms

3.1. Region Discovery Algorithm: Supervised Clustering Using Multi-Resolution Grids (SCMRG)

We have developed an algorithm called Supervised Clustering using Multi-Resolution Grids (SCMRG) [17] to identify promising regions. The SCMRG algorithm is a hierarchical grid-based method that utilizes a divisive, top-down search: each cell at a higher level is partitioned further into a number of smaller cells, and this process continues if the sum of the rewards of the lower level cells is greater than the obtained reward for the cell at the higher level. The returned cells usually have different sizes, because they were obtained at different level of resolution. A queue data structure is used to store all the cells that need be processed. The example in Figure 2 explains the procedure of this algorithm using a sample dataset, where are two regions are identified. The algorithm starts at a user defined level of resolution, and considers the following three cases when processing a cell c.

1. Case 1. If the cell c receives a reward, and its reward is greater than the sum of the rewards of its children and the sum of rewards of its grandchildren respectively, this cell is returned as a cluster by the algorithm; e.g., c_{11} in Figure 2.

2. Case 2. If the cell c does not receive a reward, nor does its children and grandchildren, neither the cell nor any of its decedents will be further process or labeled as a cluster; e.g., c_{14} in Figure 2.

3. Case 3. Otherwise, if the cell c does not receive a reward, but its children receive rewards, put all the children of the cell c into a queue for further processing. e.g., c_{13} in Figure 2.

The algorithm traverses through the hierarchical structure and examines those cells in the queue. This hierarchical grid-based approach captures clustering information associated with spatial cells without recourse to the individual objects and it does not drill down a cell if it does not look so promising (case 2). The advantage is that the computational complexity is linear with the number of grid cells processed, which is usually much less than the number of objects. Thus the algorithm is capable of processing large datasets efficiently. The employed framework has some similarity with the framework introduced in the STING algorithm [18]. The difference is that our algorithm focuses on finding interesting cells (that receive high rewards) instead of cells that contain answers to a given query. Moreover, it only computes cell statistics when needed and not in advance as STING does.

3.2. Generation of Regional Rules

Once regions are identified, we construct frequent itemsets for each region. Extending the Apriori algorithm [2] by utilizing a given class structure, our method enforces that each candidate k-itemset include at least one class label. After frequent itemsets are generated, we use the same approach proposed by the Apriori algorithm to generate strong rules using the $min\_confidence$ threshold.

4. A Real-World Case Study: Discover Pattern of Rick from Arsenic

4.1. Datasets: Data Collection and Data Preprocessing

The arsenic datasets used in this study are extracted from the Texas Ground Water Database (GWDB) maintained by the Texas Water Development Board [3]. Arsenic in very high concentrations is poisonous. Low-level, long term exposure to arsenic can lead to increased risk of cancer [7].

Because data collection and maintenance procedures and standards have been changed over the years in the GWDB, datasets have to be cleaned to deal with problems such as missing values, inconsistent data, and duplicate entries. The obtained arsenic spatial dataset includes spatial attributes (S), non-spatial attributes (A), and class labels (CL) for each water well. Some of the spatial attributes are directly extracted from the database, such as river basin, zone, latitude and longitude. Implicit spatial attributes, such as distance between wells and rivers, are estimated using the 9-intersection model [4]. Non-spatial attributes are selected with the assistance of domain experts [9, 11, 13]; they include well depth, concentration of fluoride, nitrate, and other chemical metal elements, such as vanadium, iron, molybdenum, selenium, etc. We classify water wells into two classes: “safe” and “dangerous”. Based on the standard for drinking water by Environment Protection Agency [1]: a well is considered “dangerous” if its arsenic concentration level is above $10\mu g/l$. To ensure the quality of our study, we have selected 9,939 records [1] from the original 14,358 samples. Figure 3 illustrates arsenic concentration in Texas, where safe wells are in green (or light grey), dangerous wells in red (or dark grey).

4.2. Experimental Results Evaluation

A region whose arsenic distribution is significantly higher/lower (high reward value with respect to “dangerous”/”safe”) is considered as an arsenic *hotspot*/*coldspot*. In our study, we re-discovered several hotspots and coldspots, which have been studied by geoscientists before. We are presenting our results with validation from the published results in geoscience for both regional discovery and association rule mining.

In the region discovery, the SCMRG algorithm is applied to a dataset that consists of longitude and latitude of wells along with arsenic class labels (“dangerous” or “safe”). Figure 4 depicts the result of such a run that identifies 4 subregions. Specifically, Region 1 and 3 have high density of dangerous wells, and Region 2 and 4 have high density of safe wells. Hotspot Region 1 overlaps with the arsenic risk zone reported in National Water-Quality Assessment Program [14], and hotspot Region 3 is confirmed as an arsenic risk zone by Parker’s work published in the Natural Arsenic in Groundwater [13].

In the regional association rule mining, we set $min\_support$ to 10% and $min\_confidence$ to 70%. Mining regional rules in arsenic hotspots discovers attributes that are associated with high arsenic concentrations, and in coldspots discovers attributes related with low arsenic concentrations. We present the rules for the 4 highly rewarded subregions investigated in the following, all meaningful and important according to arsenic study literature. e.g., in Region 3 of Figure 4, we discover:

$$is\_a(X, Well) \wedge nitrate(X, 0 - 0.085)$$

[1] Sampled in accordance with the Texas Water Development Board’s *A Field Manual for Ground-Water Sampling,* 1990. “Samples are collected when temperature, conductivity, and PH have stabilized. The sample is filtered and field tested for alkalinity. Samples are preserved as applicable, kept chilled, and delivered to the lab. Holding times are honored. Organic sub-samples are not filtered” [3].

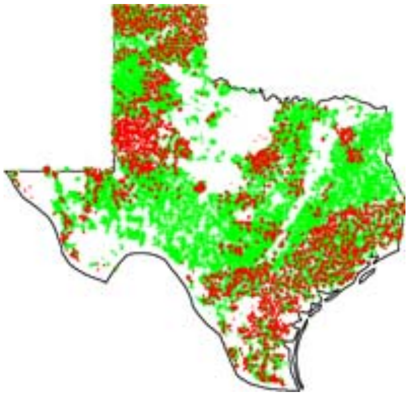

Figure 3. Map of Texas showing arsenic concentration level. Legend: green (or light grey) star – safe wells; red (or dark grey) dot – dangerous wells.

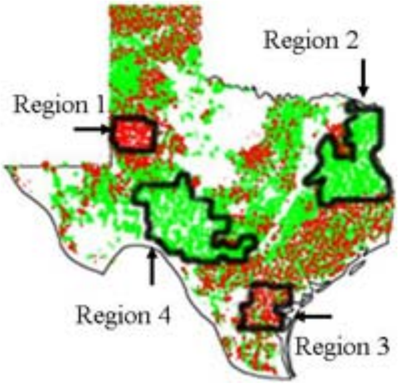

Figure 4. Interesting regions are identified using $\beta = 1.01$, $\eta = 1$, $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 1$. Average region purity = 0.85.

$$\rightarrow \mathit{aresnic\_level}(X,\ \mathit{dangerous})\ \ (100\%). \quad (1)$$

The rule states with 100% confidence that wells in Region 3 with nitrate concentration lower than $0.085 mg/l$ have dangerous arsenic concentration level. The strong association between nitrate and high arsenic concentration level is verified by Hudak's work [9] in an environmental geology study.

Our experiment results also show some novel rules that have not been analyzed in the literature of arsenic analysis; e.g., in Region 1 the following rule is discovered:

$$\begin{aligned} &\mathit{is\_a}(X,\mathit{Well}) \wedge \mathit{depth}(X, 0-215.5) \wedge \mathit{iron}(19.65-20.05) \\ &\qquad \rightarrow \mathit{aresnic\_level}(X,\ \mathit{dangerous})\ (100\%). \end{aligned} \quad (2)$$

The rule indicates that a certain range of well depth and iron concentration level are associated high arsenic concentrations. We hope that the results from our study will help the domain experts in selecting interesting hypothesis for further scientific exploration, without the need to have to analyze complex casual relationships initially.

Furthermore, we are interested to know whether the rules are different in different regions. We compared the sets of rules generated for Region 1 and Region 3 (hotspots), Region 2 and Region 4 (coldspots). The spatial risk patterns associated with arsenic are very different in each region. e.g., comparing the rule 1 identified in Region 3 with the rule 3 extracted from the Region 1:

$$\begin{aligned} &\mathit{is\_a}(X,\mathit{Well}) \wedge \mathit{nitrate}(X, 28.085-\infty) \wedge \\ &\qquad \wedge \mathit{fluoride}(X, 4.605-\infty) \\ &\rightarrow \mathit{aresnic\_level}(X,\ \mathit{dangerous})\ \ (100\%). \end{aligned} \quad (3)$$

Instead of being related with relatively low concentration of nitrate (< 0.085), the rule says that with 100% confidence, wells in Region 3, with nitrate concentration higher than 28.085 mg/l, and fluoride concentration higher than 4.605 mg/l, have dangerous arsenic concentration level.

Rules in coldspots Region 2 and 4 shed lights on what may prevent high arsenic concentrations. e.g., we find the following rule, discovered both in Region 2 and 4, states what is associated with low arsenic concentrations.

$$\begin{aligned} &\mathit{is\_a}(X,\mathit{Well}) \wedge \mathit{nitrate}(X, 0.455-16.1) \wedge \\ &\mathit{fluoride}(X, 0.095-0.315) \wedge \mathit{vanadium}(X, 3.25-5.945) \\ &\qquad \rightarrow \mathit{aresnic\_level}(X,\ \mathit{safe})\ (100\%)\ (4) \end{aligned}$$

As comparison, we also mine supervised association rules in the whole dataset. After some exploratory experiments, we found that by reducing the *min_support* from 10% to 1%, we are able to identify more interesting rules globally. However, in this case more than 100,000 rules are generated. Compared with the 300 rules on average per region in regional rule mining, it is laborsome to go through all those rules to find any meaningful ones. However, the four rules that we discovered in subregions are failed to be identified in the global level, the state of Texas. Statewide rule mining finds very general rules, such as:

$$\begin{aligned} &\mathit{is\_a}(X,\mathit{Well}) \wedge \mathit{water\_use}(X, \text{"}\mathit{by\ humam\ beings}\text{"}) \wedge \\ &\qquad \mathit{arsenic\_level}(X, \mathit{safe}) \\ &\qquad \rightarrow \mathit{inside}(X, \mathit{Basin}19)\ \ (86\%) \end{aligned} \quad (5)$$

It says that wells used by human beings, with safe arsenic concentration level are very likely (confidence is 86%) located in river basin 19.

In summary, from these experiments we identified meaningful regions at different granularity and regional rules based on our proposed framework and algorithms. We also confirmed what has been observed by researchers in geoscience, that regional rules are not the representative of global rules, and vice versa.

5. Conclusions

One critical requirement for spatial data mining is the capability to analyze datasets at different levels of granularity, in addition to analyze data globally. Furthermore, it is desirable to have the capability to move between different granularities, particularly if the obtained results are unsatisfactory. We also provided evidence that discovering regional patterns is very important in spatial data mining. Unfortunately, the currently employed association rule mining techniques do not offer such capability. We see our work as a first step toward providing such capabilities.

This paper centers on discovering regional association rules in spatial datasets. In particular, we introduce a novel framework to mine regional association rules relying on a given class structure: transaction are assumed to belong to a finite set of classes. A reward-based region discovery method has been proposed that allows identifying interesting subregions in spatial datasets for which regional association rules are then generated. In addition, a novel, divisive, grid-based supervised clustering algorithm named SCMRG has been discussed that searches for interesting regions in large spatial datasets, maximizing a reward-based fitness function that measures the interestingness of a given set of regions. Then, an integrated approach is presented to systematically mine regional rules.

We evaluated the proposed framework on a real-world case study to identify spatial risk patterns from arsenic in Texas water supply. We identified arsenic hotspots and coldspots and created regional rules from the obtained regions, rediscovering several relationships that are already reported in the scientific literature. Moreover, our approach identified several new relationships between arsenic and other factors that provide scientists with novel hypotheses that deserve further exploration in future research.

References

[1] Environmental Protection Agency. *http://www.epa.gov/*, 2006.

[2] R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, 26–28 1993.

[3] Texas Water Development Board. *http://www.twdb.state.tx.us/home/index.asp*, 2006.

[4] M. J. Egenhofer and R. D. Franzosa. Pointset topological spatial relations. *International Journal for Geographical Information Systems*, 5(2):161–174, 1991.

[5] C.F. Eick, B. Vaezian, D. Jiang, and J. Wang. Discovering of interesting regions in spatial data sets using supervised cluster,. In *PKDD'06, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006.

[6] C.F. Eick, N. Zeidat, and Z. Zhao. Supervised clustering: Algorithms and application. In *International Conference on Tools with AI*, pages 774–776, 2004.

[7] A. H. Smith et al. Cancer risks from arsenic in drinking water. In *Environmental Health Perspectives*, volume 97, pages 259–267, 1992.

[8] M. F. Goodchild. The fundamental laws of GIScience. Invited talk at University Consortium for Geographic Information Science, University of California, Santa Barbara, 2003.

[9] P. F. Hudak. Arsenic, nitrate, chloride and bromide contamination in the gulf coast aquifer, south-central Texas, USA. *International Journal of Environmental Studies*, 60:123–133, 2003.

[10] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In *Proc. 4th Int. Symp. Advances in Spatial Databases, SSD*, volume 951, pages 47–66, 6–9 1995.

[11] L. M. Lee and B. Herbert. A GIS survey of arsenic and other trace metals in groundwater resources of Texas. In *Natural Arsenic in Groundwater: Science, Regulation, and Health Implications*, 2001.

[12] S. Openshaw. Geographical data mining: Key design issues. In *GeoComputation*, 1999.

[13] R. Parker. Ground water discharge from mid-tertiary rhyolitic ash-rich sediments as the source of elevated arsenic in south texas surface waters. In *Natural Arsenic in Groundwater: Science, Regulation, and Health Implications*, 2001.

[14] National Water Quality Assessment Program. Ground-water quality of the southern high plains aquifer, Texas and New Mexico. Technical report, U.S. Department of the Interior and U.S. Geological Survey, 2001.

[15] S. Shekhar. Spatial data mining: Accomplishments and research needs. Keynote speech at GIScience 2004, 2004.

[16] S. Shekhar and S. Chawla. *Spatial Databases: A Tour*. Prentice Hall, 2003 (ISBN 013-017480-7), 2003.

[17] J. Wang. Region discovery using hierarchical supervised clustering. Master's thesis, Computer Science Department, Univeristy of Houston, May 2006.

[18] W. Wang, J. Y., and R. R. Muntz. STING: A statistical information grid approach to spatial data mining. In *Twenty-Third International Conference on Very Large Data Bases*, pages 186–195. Morgan Kaufmann, 1997.