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 of†

Abstract

The immense explosion of geographically referenced data calls for efficient discovery of spatial knowledge. One critical requirement for spatial data mining is the capability to analyze datasets at different levels of granularity. 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. Unfortunately, most of the current data mining techniques are ill-prepared for discovering regional knowledge. For example, when using traditional association rule mining, regional patterns frequently fail to be discovered due to insufficient global confidence and/or support. This raises the questions on how to measure the interestingness of a set of regions and how to search effectively and efficiently for interesting regions. 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 Texas water supply.

1. Introduction

Advanced data collecting tools in digital mapping, remote sensing, and global diffusion of Geographic Information Systems (GIS) are generating increasingly large spatial datasets. NASA's Earth Observing System (EOS) [28] has offered long-term global observation of the land surface, biosphere, solid Earth, atmosphere, and ocean in the rate of a terabyte of data per day since 1999. Furthermore, more powerful and reliable location-enabled mobile devices are generating large spatial datasets. These spatial datasets contain nuggets of valuable information that call for efficient discovery of spatial knowledge. The goal of spatial data mining is to automate the extraction of interesting, useful but implicit spatial patterns [8, 13, 15, 16, 19, 25, 29, 31, 32, 33, 36].

Although data mining has been recognized as a key means of finding patterns in large datasets, traditional data mining methods alone are not sufficient for spatial data mining. 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. It has been pointed out in literature [12, 23, 30] 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 [12, 21, 22].

Unfortunately, most of the current data mining techniques are ill-prepared for discovering regional knowledge. For example, when using traditional association rule mining, 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. One frequently chosen approach is to select regions to be mined based on a priori given structure, such as a grid structure based on longitude and latitude or political boundaries; for example, using counties as subregions of a state. Unfortunately, the surface boundary of the so constructed regions frequently does not match the surface boundary of the interesting patterns, making them unlikely to be discovered. For example, let us assume that there are high arsenic con-

*Also, Computer Science Department, University of Houston-Clear Lake, Texas

†$Revision: 2545$

Table 1. A two-way contingency table between the well depth and arsenic concentration.

Well Depth	**Arsenic** dangerous	**Concentration** safe	**Total**
(0, 215.5]	1000	1000	2000
(215.5, ∞)	1200	800	2000
	2200	1800	4000

Table 2. A three-way contingency table between geographic zone A and zone B.

	Well Depth	**Arsenic** dangerous	**Concentration** safe	**Total**
ZoneA	(0, 215.5]	400	100	500
	(215.5, ∞)	1050	450	1500
ZoneB	(0, 215.5]	600	900	1500
	(215.5, ∞)	150	350	500
		2200	1800	4000

centrations along a river that crosses multiple counties in Texas. In this case, mining regional rules at county level is unlikely to detect this pattern, due to Simpson's paradox (see Section 1.1).

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 1.1 discusses Simpson's paradox. Section 1.2 reviews related work. 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.

1.1. Simpson's Paradox

A well known issue in spatial datasets is that global patterns can be very different from regional patterns. This phenomenon is known as Simpson's paradox [7], or spatial heterogeneity [30]. We illustrate the nature of this paradox using table 1 and 2.

Consider the relationship between well depth and arsenic concentration as shown in Table 1. The following rule suggests that a well up to 215.5-feet deep is associated with dangerous arsenic levels.

$$sample\_rule: \quad is\_a(X, well) \wedge depth(X, 0-215.5) \rightarrow arsenic\_level(danderous).$$

However, whether the rule holds or not depend on where the well locates. Let's assume that the minimum confidence threshold is 70%. We now calculate the confidence of the *sample_rule* globally (zone A and Zone B in table 1) and locally (Zone A and Zone B separately in table 2). In table 1, the rule is not strong enough to be identified globally because its confidence value 50% is less than the 70% minimum confidence threshold:

$$\textbf{confidence}(sample\_rule) \quad = \quad 1000/2000 = 50\%.$$

While in Zone A (Table 2) the rule holds because its confidence value 80% is above the 70% minimum confidence:

$$\textbf{confidence}(sample\_rule) = 400/500 = 80\%.$$

This rule does not hold in zone B (Table 2):

$$\textbf{confidence}(sample\_rule) = 600/1500 = 40\%.$$

Hence a well up to 215.5-feet deep is *positively associated* with dangerous arsenic levels in zone A but is *negatively associated* in the combined dataset. The reversal in the direction of association is known as Simpson's paradox.

The Simpson's paradox occurs when some underlying variables (in our example, zones) that have a large effect on the ratios of stratified data (in our example, Zone A and Zone B are two strata). Proper stratification proves to be an effective way to avoid generating spurious patterns resulting from Simpson's paradox [20]. In this paper, we aggregate spatial objects using a given measure of interestingness to identify regions. Then on those regions we mine regional association rules, which might not be discovered in a global search.

1.2. Related Work

The areas most relevant to our work are: spatial association rule mining, co-location rule discovery, supervised and semi-supervised clustering.

Association rule mining has been introduced in [2] to mine interesting relationships hidden in market basket transactions. Spatial association rule mining [16] extends association rule mining to spatial datasets. A spatial association rule takes the form of

$$P_1 \wedge P_2 \wedge ... \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge ... \wedge Q_n \quad (sup\%, \ con\%).$$

It denotes association relationships among a set of predicates P_i $(i = 1, ..., m)$ and Q_j $(j = 1, ..., n)$, where there

exists at least one spatial predicate. Spatial predicates may represent topological relationships between spatial objects (e.g., intersects, contains), or indicate a spatial orientation (e.g., north, left). The support of the rule (sup%), measures the percentage of transactions containing both the antecedent and consequent of the rule. The confidence of the rule (con%) indicates that con% of transactions that satisfy the antecedent of the rule will also satisfy the consequent of the rule. A rule $P_1 \wedge P_2 \wedge ... \wedge P_m \rightarrow Q_1 \wedge Q_2 \wedge ... \wedge Q_n$ is *strong* if sup% and con% satisfy minimum support and minimum confidence thresholds.

A common strategy used in spatial association rule mining is to decompose the problem into three subtasks:

1. Item representation and transaction definition: define "items" and "transactions" from spatial datasets.
2. Frequent itemset generation: find all the itemsets that satisfy the minimum support threshold.
3. Rule generation: construct rules from the frequent itemsets that satisfy the minimum confidence threshold.

Apriori-style [2] association mining algorithms require that objects are described using nominal attributes. Therefore continuous attributes have to be transformed into appropriate formats. In addition, transaction definition is implicit in spatial space. If spatial association rule discovery is restricted to a reference feature (such as cities or wells), then transactions can be defined using the instances of this reference feature, as in [16]. Otherwise, transactions must be generated by mining algorithms, as in spatial co-location mining by [32]. This paper adopts the transaction model in [16]. While co-location rule discovery identifies subsets of spatial features frequently located together [32]. It focuses on finding frequent, global patterns that characterize the complete dataset, whereas our approach centers on discovering regional patterns.

Supervised clustering [8, 9] focuses on partitioning classified examples, maximizing cluster purity while keeping the number of clusters low. Semi-supervised clustering employs a small amount of labeled data to aid unsupervised learning [3]. This paper applies supervised clustering to a new problem: region discovery in spatial datasets containing classified examples, then association rule mining is performed in the obtained regions.

2. Problem Formulation and Integrated Framework for Regional Association Rule Mining

There are two phases in the proposed integrated framework for regional association rule mining (Figure 1):

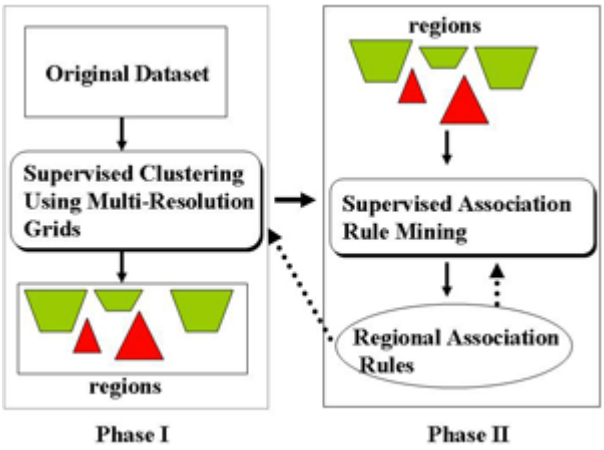

Figure 1. The Integrated Framework for Regional Association Rule Mining.

1. Phase I: Discover and identify interesting regions.
2. Phase II: Spatial association rule mining for each identified region.

In the first phase, 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 and how to identify regions using a given measure of interestingness. In the second phase, the regions are considered one at a time and all frequent itemsets for that region are generated. Regional association rules are then constructed 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 leads to 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 confidence *con* and support *sup* where

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

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 and nomenclature, 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

The first challenge in the region discovery is how the interestingness is measured. In this section, we give formal definition of a region and its interestingness measurement.

A region is 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. We find subregions $R_1,...,R_m$ such that:

1. $EXT(R_i) \subset EXT(R)$.
2. The subregions are disjoint: $EXT(R_i) \cap EXT(R_j) = \emptyset$, $i \neq j$.
3. The spatial objects $EXT(R_i)$ in region R_i maximize a given measure of interestingness. For example, one measure could be the purity of region – most or all objects in R_i belong to the same class, which is equivalent to $EXT(R_i)$ having a very low entropy with respect to the underlying class structure CL.
4. The generated regions are not required to be exhaustive with respect to R, that is, $EXT(R_1) \cup ... \cup EXT(R_m) \subseteq EXT(R)$.

Our region discovery algorithm employs a reward-based evaluation scheme that evaluates the quality of the generated regions. The fitness function that evaluates the quality of the generated regions $R_X = \{EXT(R_1),...,EXT(R_m)\}$ is defined as the sum of the rewards obtained from each region R_i $(i = 1..m)$ (Equation 1).

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

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.

2.3. Reward Function for Regional Association Rule Mining

Different reward functions that correspond to various domain interest can easily be supported in this framework. For example, a reward function can be designed to find regions that favor the discovery of co-location patterns described in [32]. Or we can use a reward function to facilitate class association rule mining [18, 37] to build more accurate regional classifiers.

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 [8]. The measure is based on a class set of class labels CL. It rewards regions in which the probability distribution of CL significantly deviates from its prior probability relying on a reward function τ. A region is a *hotspot* if its probability distribution of CL is significantly higher than the expected probability. A region is a *coldspot* if its probability distribution of CL is significantly lower than the expected probability.

Let N denotes number of objects in a dataset $\mathbb{D}$, x_i the ith 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

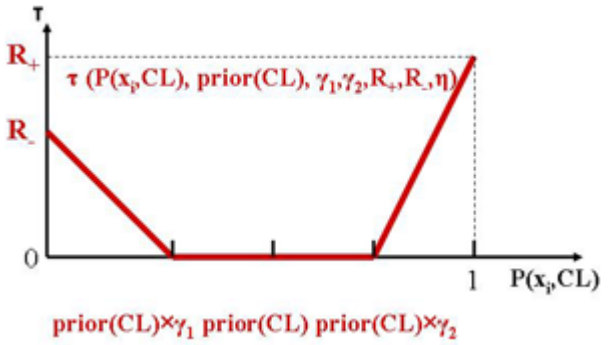

Figure 2. The reward function τ when $\eta = 1$

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 reward function 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 $\mathbb{D}$ belonging to the CL. R_+ and R_- are the maximum reward 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 2. 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 cluster.

Example 1 explains how to calculate the fitness of a clustering schema X of a sample dataset in Figure 3.

Example 1. Let us assume a clustering schema X is evaluated with respect to a class of interest "dangerous" (high-level arsenic concentrations) with $prior(dangerous) = 0.2$ and a dataset that contains 1000 examples. Suppose that the dataset is subdivided into 4 clusters $X = \{x_{11}, x_{12}, x_{13}, x_{14}\}$ at level 1, and $|x_{11}| = 50$, $|x_{12}| = 200$, $|x_{13}| = 400$, $|x_{14}| = 350$. Assume that there are 20, 100, 80, and 0 objects labeled with "dangerous" class in the 4 clusters respectively. $P(x_{11}, dangerous) = \frac{20}{50} = 0.4$, $P(x_{12}, dangerous) = \frac{100}{200} = 0.5$, $P(x_{13}, dangerous) = \frac{80}{400} = 0.2$, $P(x_{14}, dangerous) = \frac{0}{350} = 0$. The parameters used in the fitness function are as follows: $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 0$. Hence $prior(CL) \times \gamma_1 = 0.2 \times 0.5 = 0.1$, $prior(CL) \times \gamma_2 = 0.2 \times 1.5 = 0.3$. With this setting, a cluster does not receive any reward if its probability of class "dangerous" are in the range of $[0,\ 0.1]$ (due to $R- = 0$) and $[0.1,\ 0.3]$ (due to the values of $prior(CL) \times \gamma_1$ and $prior(CL) \times \gamma_2$). Therefore, no reward is given to cluster x_{13} and x_{14}. The reward for the remaining clusters are

$$\tau(x_{11}) = (\frac{0.4 - 0.3}{1 - 0.3})^1 = \frac{1}{7},$$
$$\tau(x_{12}) = (\frac{0.5 - 0.3}{1 - 0.3})^1 = \frac{2}{7}.$$

The fitness value of the clustering schema X is

$$\begin{aligned} q(X) = & \ \tfrac{1}{7} \times (\tfrac{50}{1000})^{1.1} + \tfrac{2}{7} \times (\tfrac{200}{1000})^{1.1} \ + \\ & 0 \times (\tfrac{400}{1000})^{1.1} + 0 \times (\tfrac{350}{1000})^{1.1} \\ = & \ 0.012 \end{aligned}$$

3. Algorithms

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

In this section, we first give details of our region discovery algorithm based on the reward and fitness function defined in section 2. Then we explain our regional rule generation algorithm.

We have developed an algorithm called Supervised Clustering using Multi-Resolution Grids (SCMRG) [35] 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 in the lower level, 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 algorithm (see Algorithm 1) starts at a user defined level of resolution, and considers the following three cases when processing a cell c.

Algorithm 1 The Algorithm of Supervised Clustering using Multi-Resolution Grids (SCMRG).

SCMRG (*min_cell_size*)
1. Determine a level of resolution l to start with.
2. Assign spatial objects to grid cells.
3. **for** each cell c at current level l **do**
4. enqueue(c, *cellQueue*).
5. **end for**
6. **while** *NOT empty*(*cellQueue*) **do**
7. c= dequeue(*cellQueue*).
8. r = reward (c). {Calculate reward for the cell.}
9. **for** each $c_{child} \in$succ(c) **do**
10. $r_{children} = r_{children}$+reward ($c_{child}$).
11. **end for** {Calculate reward for its children.}
12. **for** each $c_{grandchild} \in$succ(succ(c)) **do**
13. $r_{grandchildren} = r_{grandchildren}$+reward ($c_{grandchild}$).
14. **end for** {Calculate reward for its grandchildren.}
15. **if** $r > 0$ {The cell receives a reward.}
16. **if** $r > r_{children}$ *AND* $r > r_{grandchildren}$
17. label the cell as a cluster.
18. **else** {The cell should be divided further.}
19. **if** (the size of each $c_{child} \in$succ(c) $>$*min_cell_size*)
20. enqueue(*succ*(c), *cellQueue*).
21. **end if**
22. **end if**
23. **else if** $r = 0$ {The cell does not receive a reward.)
24. **if** *NOT* ($r_{children} = 0$ *AND* $r_{grandchildren} = 0$)
25. **if** (the size of each $c_{child} \in$succ(c) $>$*min_cell_size*)
26. enqueue(*succ*(c), *cellQueue*).
27. **end if**
28. **end if** {The cell should be divided further}
29. **end if**
30. **end while**
31. Collect all the cluster-labeled cells from different levels.
32. Obtain regions by merging neighbor clusters if it improves the fitness.
33. Return the obtained regions.

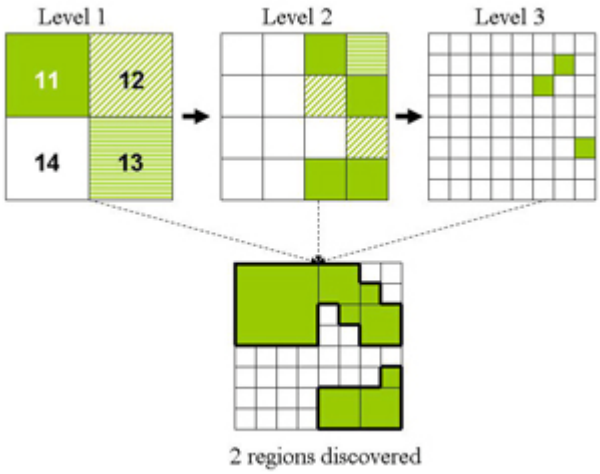

Figure 3. Running the SCMRG algorithm on a sample dataset.

1. Case 1. If the cell c receives a reward, and its reward is greater than the sum of the rewards of its children (*succ*(c)) and greater than the sum of rewards of its grandchildren, this cell is returned as a cluster by the algorithm (step 15-17).

2. Case 2. If the cell c does not receive a reward and its children and grandchildren do not receive a reward, neither the cell nor any of its decedents will be labeled as a cluster (step 23-29).

3. Case 3. Otherwise, put all the children of the cell c (*succ*(c)) into a queue for further processing (step 18-21, step 24-28).

We traverse through the hierarchical structure and exam those cells in the queue from the higher level. The algorithm uses a user-defined cell size as a depth bound. Cells that are smaller than this cell size will not be split any further (step 19, step 25). Finally, we collect all the cells that have been identified in case 1 from different levels, and we merge neighbor clusters if it improves the fitness as defined in equation 2. The obtained regions are returned as the result of executing SCMRG (step 31-33).

This hierarchical grid-based approach captures clustering information associated with spatial cells without recourse to the individual objects as we do 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 [36]. The difference is that

our algorithm focuses on finding interesting cells (that receive high reward) instead of cells that contain answers to a given query. And it only computes cell statistics when needed and not in advance as STING does.

The example in Figure 3 explains the procedure of this algorithm using a sample dataset. The first decomposition results in 4 cells $c_{11}, c_{12}, c_{13}, c_{14}$ at level 1. As illustrated in Example 1, only c_{11} and c_{12} receive rewards. Assume the reward of c_{11} is greater than the sum of the rewards of its children and greater than the sum of rewards of its grandchildren, c_{11} is then labeled as a cluster according to case 1. The cell, c_{14}, does not receive any rewards. Assume neither do its children nor grandchildren receive any rewards. According to case 2, c_{14} is not labeled as a cluster, and its successors are not saved into the queue. Although the cell, c_{13}, receives no reward, assume its children receive rewards. According to case 3, all the children of c_{13} are saved into a queue to be further processed. The cells at level 1 are then divided into level 2 and 3 and the same procedure is applied to all the cells in the queue. Each cell is labeled accordingly. The intermediate results are shown at level 2 and 3 in Figure 3. Neighbor clusters are then merged if it improves the fitness. In this example, there are two regions identified.

3.2. Generation of Regional Rules

Once regions are identified, we construct frequent itemsets for each region. Our Supervised_Apriori_Gen algorithm (see Algorithm 2) extends the Apriori algorithm [2] by utilizing a given class structure.

The Apriori algorithm first makes a single pass over the data set to determine the support of each single item, which generates all frequent 1-itemsets, F_1. Next, the algorithm iteratively generates candidate k-itemsets using the frequent (k-1)-itemsets found in the previous iteration. Candidate itemsets are pruned if it is not frequent. The algorithm terminates when there are no new frequent itemsets generated, e.g., F_k=Ø. The given class structure is incorporated in our Supervised_Apriori_Gen algorithm by enforcing that each candidate k-itemset must include at least one class label; otherwise it is pruned even it is frequent. The Supervised-Apriori-Gen uses the $F_{k-1} \times F_{k-1}$ method [34] to merge a pair of frequent (k-2)-itemset . Basically, let $A = \{a_1, a_2, ..., a_{k-1}\}$ and $B = \{b_1, b_2, ..., b_{k-1}\}$ be a pair of frequent (k-1)-itemset. A and B are merged if they satisfy the following conditions:

$$a_i = b_i \quad (for\ i = 1, 2, ..., k-2)\ \ and\ \ a_{k-1} \neq b_{k-1}$$

Supervised-Apriori-Gen algorithm initially starts with candidate 2-itemset construction, which is the base of the K-itemset generation ($k > 2$). To ensure each 2-itemset

Algorithm 2 Candidate Generation and Pruning: **Supervised_Apriori_Gen**

Supervised_Apriori_Gen(F_{k-1})
1. **if** $k = 2$ {Deal with candidate 1- and 2-itemsets}
2. **for** each frequent 1-itemset $f \in F_1$ **do**
3. insert f into C_1. {Generate candidate 1-itemsets}
4. **end for**
5. $(C_{1\_class\_label}, C_{1\_other}) = split(C_1, CL)$.
{Split C_1 , group class labels into $C_{1\_class\_label}$, and the other frequent 1-itemsets into C_{1_other}}.
6. **for** each candidate itemset $c1 \in C_{1\_label}$ **do** {Generate candidate 2-itemsets with class-label items and non-class-label items}
7. **for** each candidate itemset $c2 \in C_{1\_other}$ **do**
8. c = form $c1$ and $c2$.
9. insert c into C_2. {Generate candidate 2-itemsets}
10. **end for**
11. **end for**
12. **for** each candidate itemset $c1 \in C_{1\_label}$ **do**
13. $C_{post} = subset\_split(C_{1\_label}, c1)$. {Identify all the class labels in the array C_{1_label} that is located after $c1$}
14. **for** each candidate itemset $c2 \in C_{post}$) **do**
15. c = form $c1$ and $c2$.
16. insert c into C_2.
17. **end for**
18. **end for**
19. **else**
20. **for** each $i1$ in F_{k-1}
21. **for** each $i2$ in F_{k-1}
22. **if** (first $k-2$ items of $i1, i2$ same) $\wedge$(last item of $i1, i2$ differs)
23. c =form (first $k-1$ items of $i1$) and (last item of $i2$).
24. insert c into C_k
25. **end if**
26. **end for**
27. **end for**
28. **end if**
29. return C_k

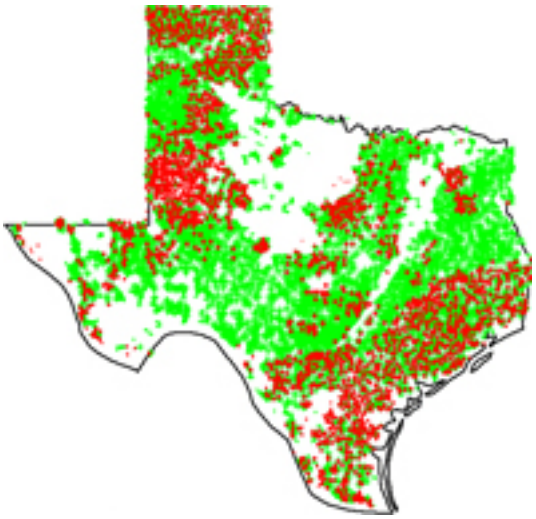

Figure 4. Map of Texas showing arsenic concentration level. Legend: green (or light grey) star – safe wells; red (or dark grey) dot – dangerous wells.

must include at least one class label, the algorithm first constructs candidate 1-itemsets from frequent 1-itemset (step 2-4). Second, to generate candidate 2-itemsets with class labels, the algorithm separates class-label items from other items with *split* function (step 5). Then the algorithm enumerates class-label items with the rest items (step 6-11), as well as class-label items with themselves (step 12-18). Thus step 6-11 generate candidate 2-itemsets formed between class labels and other non-class-label items; step 12-18 generate candidate 2-itemsets formed between class labels. The 2-itemsets are then used for K-itemsets generation ($K > 2$) (step 19-26).

After frequent itemsets are generated, we use the same approach proposed by the Apriori algorithm to generate strong supervised rules using *min_confidence* threshold.

4. A Real-World Case Study: Arsenic Spatial Risk Pattern Discovery in Texas

In this section we describe the experiment procedures of applying the proposed framework on a real world case study that identifies arsenic spatial risk patterns in Texas water supply. Then we present and discuss the experimental results and evaluate the performance of the proposed framework.

The experiments are conducted in four steps:

1. Data collection and data preprocessing, including data cleaning, transforming continuous attributes into nominal attributes, and constructing transactions using water well as the reference feature.
2. Identifying arsenic *coldspots* and *hotspots*. In this paper, a region whose arsenic distribution is significantly higher (high reward value with respect to "dangerous") is considered as an arsenic hotspot; a region whole arsenic distribution is significantly lower (high reward value with respect to "safe") is considered as an arsenic coldspot.
3. Mining supervised association rules from each identified region and for the complete dataset.
4. Analyzing the obtained rules and rule sets, adjust the parameter settings, and re-run the experiments in step 2 and 3 according if results are unsatisfactory.

4.1. Datasets: Data Collection and Data Preprocessing

The datasets used in this study are extracted from the Texas Ground Water Database (GWDB) maintained by the Texas Water Development Board, the state agency in charge of statewide water planning [4]. The Texas Water Board has monitored and analyzed the concentrations of this super toxic element over the last 25 years. Arsenic in very high concentrations is poisonous. Low-level, long term exposure to arsenic can lead to increased risk of cancer [10]. Arsenic is derived from both anthropogenic sources – such as drainage from mines and mine tailings, pesticides, and biocides, and from natural sources – such as hydrothermal leaching of arsenic containing minerals or rocks. The World Health Organization has reported arsenic in drinking water in U.S., Thailand, Mexico, India, Hungary, Ghana, Chile, China, Bangladesh, and Argentina [24].

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 [6]. Non-spatial attributes are selected with the assistance of domain experts [14, 17, 26]; they include well depth, concentration of fluoride, nitrate, and other chemical metal elements, such as vanadium, iron, molybdenum and 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 quality of the association rule generated in the study,

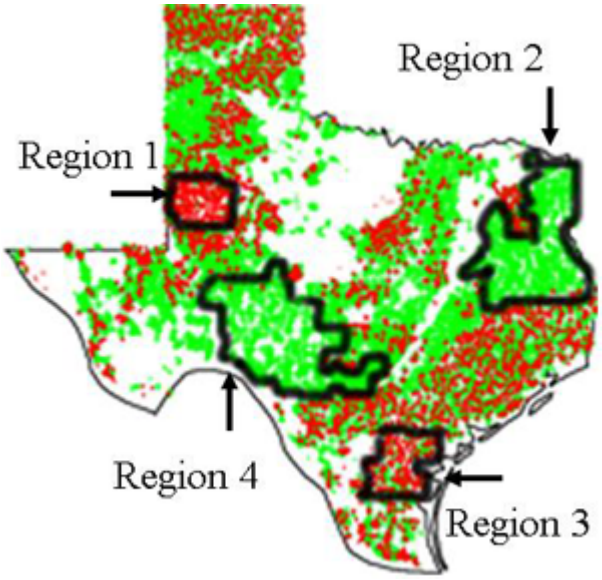

Figure 5. Interesting regions are identified using$\beta = 1.01,\ \eta = 1,\ \gamma_1 = 0.5,\ \gamma_2 = 1.5,\ R+ = 1,\ R- = 1$. Average region purity = 0.85.

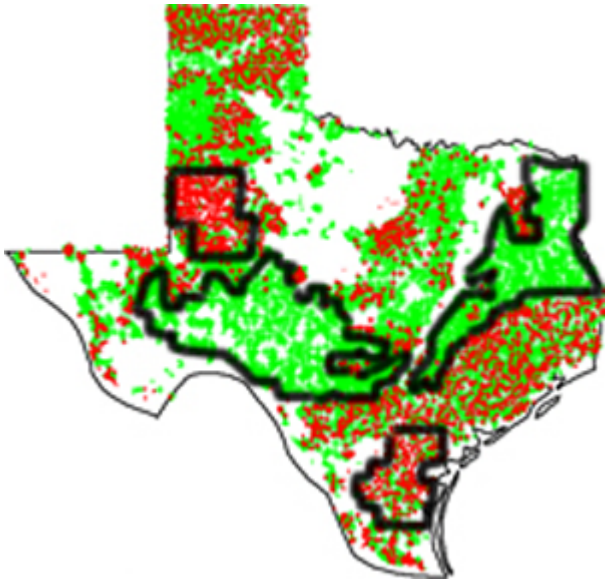

Figure 6. Interesting regions are identified using $\beta = 1.035,\ \eta = 1,\ \gamma_1 = 0.5,\ \gamma_2 = 1.5,\ R+ = 1,\ R- = 1$. Average region purity = 0.83.

we only select lab test results that used honored sampling procedures, which result in 11,922 records selected from the GWDB after the process of data cleaning. Figure 4 illustrates arsenic concentration in Texas, where safe wells are in green (or light grey), dangerous wells in red (or dark grey).

In preparation of the association rule mining, continuous attributes excluding latitude and longitude are first converted into nominal attributes using the supervised method Recursive Minimal Entropy Partitioning [11]. The supervised entropy based method uses class labels "dangerous" and "safe" to place the splits in a way that maximizes the purity of the intervals. The method usually results unequal bin size and has been proved to produce better results in data mining tasks [5]. For example, the value of nitrate concentration has been discretized into five intervals of (0,0.085], (0.085,0.455], (0.455,16.1], (16.1,28.085], (28.085,∞) (measurement unit mg/l).

4.2. Experimental Results Evaluation

From our study, we have re-discovered several interesting risk regions with high arsenic concentrations (hotspots), which have been studied by geoscientists before. We have also identified regions with low arsenic concentrations (coldspots). The association rules that we constructed from those identified regions will help the geoscientists to identify the cause of high arsenic concentration in different regions, and the common situations for those regions with low arsenic concentrations. 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 5 depicts the result of such a run that identifies four regions. 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 [27], and hotspot Region 3 is confirmed as an arsenic risk zone by Parker's work published in the Natural Arsenic in Groundwater [26].

If we are interested in finding larger regions with likely lower purity, using a larger value of β results in bigger size of regions. Figure 6 shows enlarged regions when β is increased from 1.01 to 1.035. In our experiments, we adjusted the granularity of regions by the quality of rules discovered in step 3. We observed that $\beta = 1.01,\ \eta = 1$ give us best results in the rules constructed in the supervised association rule mining.

The Supervised_Apriori_Gen algorithm is used to generate frequent itemsets for all the regions identified. We use $min\_support = 10\%$ and $min\_confidence = 70\%$ threshold for the experiments. We present the first few rules for the regions investigated, all meaningful and important according to arsenic study literature.

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. For example, in Region 3 of Figure 5, we discover:

$$is\_a(X, Well) \wedge nitrate(X, 0 - 0.085) \rightarrow aresnic\_level(X, dangerous)\ (100\%). \quad (1)$$

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 [14] in environmental geology study.

In region 1 of figure 5, we also discovered:

$$is\_a(X, Well) \wedge vanadium(X, 20.05 - 37.95) \wedge selenium(74.55 - \infty) \rightarrow aresnic\_level(X, dangerous)\ (100\%). \quad (2)$$

The rule states with 100% confidence that wells in Region 1 with vanadium concentration between 20.05 and $37.95 \mu g/l$ and selenium concentration larger than $74.55 \mu g/l$ have dangerous arsenic concentration level. Our discovery is also confirmed by Lee et al. in [17].

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

$$is\_a(X, Well) \wedge depth(X, 0 - 215.5) \wedge iron(19.65 - 20.05) \rightarrow aresnic\_level(X, dangerous)\ (100\%). \quad (3)$$

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. For example, comparing the rule 1 identified in Region 3 with the rule 4 extracted from the Region 1:

$$is\_a(X, Well) \wedge nitrate(X, 28.085 - \infty) \wedge \wedge fluoride(X, 4.605 - \infty) \rightarrow aresnic\_level(X, dangerous)\ (100\%). \quad (4)$$

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. For example, we find the following rule, discovered both in Region 2 and 4, states what is associated with low arsenic concentrations.

$$is\_a(X, Well) \wedge nitrate(X, 0.455 - 16.1) \wedge fluoride(X, 0.095 - 0.315) \wedge vanadium(X, 3.25 - 5.945) \rightarrow aresnic\_level(X, safe)\ (100\%)\ (5)$$

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 cumbersome to go through all those rules to find any meaningful ones. The need to use low support values for complete datasets has also been observed by [18]. However, all the regional rules (rule 1 to 5) that we discussed previously are not generated because they do not have enough confidence or support globally. Statewide rule mining finds very general rules, such as:

$$is\_a(X, Well) \wedge water\_use(X, \text{"by humam beings"}) \wedge arsenic\_level(X, safe) \rightarrow inside(X, Basin19)\ (86\%) \quad (6)$$

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 spatial data mining and geoscience, that regional rules are not the representative of global rules, and vise 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 of 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 hypothesis that deserve further exploration in future research.

Our future work will center on applying our techniques to larger arsenic datasets that also include population, geological, and agricultural data. We plan to construct a general framework for interpreting and evaluating risks in environmental data, such as discovery co-location patterns and ground level ozone forecasting. Finally, we plan to extend our framework to support region discovery for spatio-temporal datasets.

6. Acknowledgements

We thank Dr. Shuhab Khan (Geosciences Department, University of Houston) for his valuable comments on experiments. We thank Radu Boghici and Roger M. Quincy (Texas Water Development Board) for help on Texas Ground Water Database. We thank Dr. Ping Chen (Computer Science Department, University of Houston-Downtown) for useful discussions.

References

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

[2] Rakesh Agrawal, Tomasz Imielinski, and Arun 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, Washington, D.C., 26–28 1993.

[3] Sugato Basu, Mikhail Bilenko, and Raymond J. Mooney. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering. In *the ICML-2003 Workshop on the Continuum from Labeled to Unlabeled Data in Machine Learning and Data Mining Systems*, pages 42–49, Washington DC, August 2003.

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

[5] James Dougherty, Ron Kohavi, and Mehran Sahami. Supervised and unsupervised discretization of continuous features. In *International Conference on Machine Learning*, pages 194–202, 1995.

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

[7] Simpson EH. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, B13:238–241, 1951.

[8] Christoph Eick, Banafsheh Vaezian, Dan Jiang, and Jing Wang. Discovering of interesting regions in spatial data sets using supervised cluster, conditionally accepted for publication. In *PKDD'06, 10th European Conference on Principles and Practice of Knowledge Discovery in Databases*, 2006.

[9] Christoph F. Eick, Nidal Zeidat, and Zhenghong Zhao. Supervised clustering: Algorithms and application. In *International Conference on Tools with AI*, pages 774–776, Boca Raton, Florida, 2004.

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

[11] Usama M. Fayyad and Keki B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In Morgan Kaufmann, editor, *Proceedings of the 13th International Joint Conference on Artificial Intelligence*, pages 1022–1027, 1993.

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

[13] Jiawei Han, Micheline Kamber, and Anthony K. H. Tung. Spatial clustering methods in data mining: A survey. In *Geographic Data Mining and Knowledge Discovery*, 2001.

[14] Paul 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.

[15] Vandana Janeja and Vijayalakshmi Atluri. FS3 : A random walk based free-form spatial scan statistic for anomalous window detection. In *Fifth IEEE International Conference on Data Mining*, 2005.

[16] Krzysztof Koperski and Jiawei Han. Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer and J. R. Herring, editors, *Proc. 4th Int.*

Symp. Advances in Spatial Databases, SSD, volume 951, pages 47–66, 6–9 1995.

[17] Lai Man Lee and Bruce 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 (Posters)*, 2001.

[18] Wenmin Li, Jiawei Han, and Jian Pei. CMAR: Accurate and efficient classification based on multiple class-association rules. In *International Conference on Data Mining (ICDM'01)*, San Jose, CA, Nov. 2001.

[19] Robert Munro, Sanjay Chawla, and Pei Sun. Complex spatial relationships. In *The Third IEEE International Conference on Data Mining (ICDM2003)*, 2003.

[20] Mantel N. and Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719–748, 1959.

[21] Stan Openshaw. Two exploratory space-time attribute pattern analysers relevant to GIS. In S. Fotheringham and P. Rogerson, editors, *Spatial Analysis and GIS*, pages 83–104, London, 1994. Taylor and Francis.

[22] Stan Openshaw. Developing automated and smart spatial pattern exploration tools for geographical information systems applications. *The Statistician*, 44(1):3–16, 1995.

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

[24] World Health Organization. *http://www.who.int/*, 2006.

[25] S. Papadimitriou, A. Gionis, P. Tsaparas, A. Väisänen, H. Mannila, and C. Faloutsos. Parameter-free spatial data mining using MDL. In *5th International Conference on Data Mining (ICDM) 2005*, 2005.

[26] Ronald 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.

[27] National Water-Quality Assessment Program. Groundwater quality of the southern high plains aquifer, Texas and New Mexico, open-file report 03-345. Technical report, U.S. Department of the Interior and U.S. Geological Survey, 2001.

[28] NASA's Earth Observing System Project. *http://eospso.gsfc.nasa.gov/*, 2006.

[29] J. F. Roddick and M. Spiliopoulou. A bibliography of temporal, spatial and spatio-temporal data mining research. In *SIGKDD Explorations*, volume 1, pages 34–38, 1999.

[30] Shashi Shekhar. Spatial data mining: Accomplishments and research needs. Keynote speech at GIScience 2004 (3rd Biannual International Conference on Geographic Information Science), 2004.

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

[32] Shashi Shekhar and Yan Huang. Discovering spatial co-location patterns: A summary of results. *Lecture Notes in Computer Science*, 2121:236+, 2001.

[33] Shashi Shekhar, Pusheng Zhang, Yan Huang, and Ranga Raju Vatsavai. Book chapter in data mining: Next generation challenges and future directions. In Hillol Kargupta and Anupam Joshi, editors, *Spatial Data Mining*. AAAI/MIT Press, 2003.

[34] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining*. Addison-Wesley, 2006.

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

[36] Wei Wang, Jiong Yang, and Richard 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, Athens, Greece, 1997. Morgan Kaufmann.

[37] Xiaoxin Yin and Jiawei Han. CPAR: Classification based on predictive association rules. In *3rd SIAM International Conference on Data Mining (SDM'03)*, San Francisco, CA, May 2003.