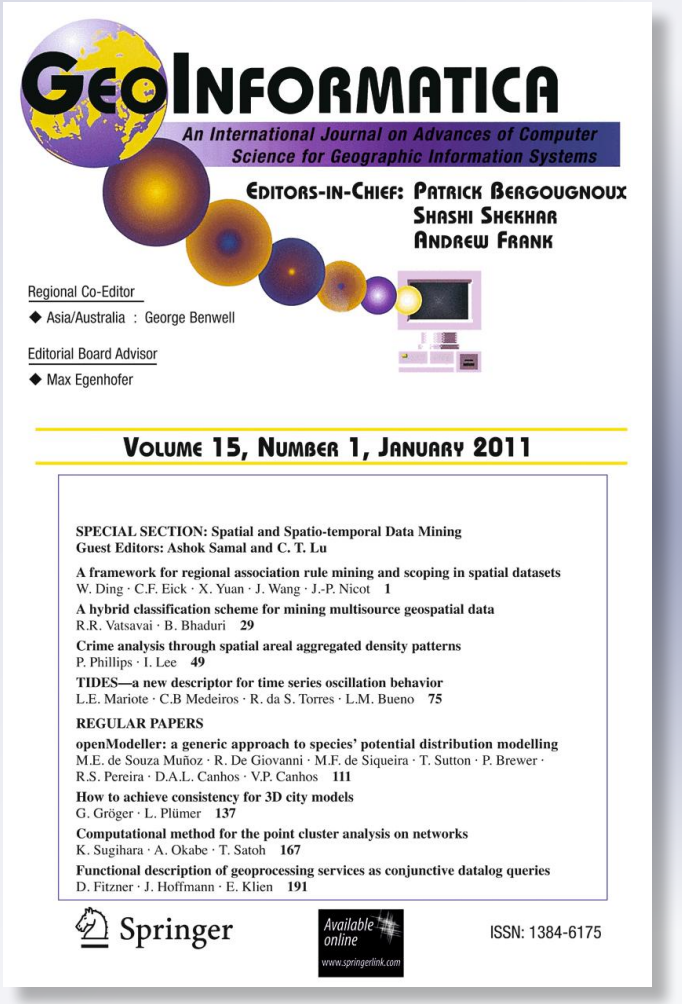
A framework for regional association rule mining and scoping in spatial datasets

GeoInformatica
An International Journal on Advances of Computer Science for Geographic Information Systems

ISSN 1384-6175
Volume 15
Number 1

Geoinformatica (2010) 15:1-28
DOI 10.1007/
s10707-010-0111-6



Springer

Geoinformatica (2011) 15:1–28
DOI 10.1007/s10707-010-0111-6

A framework for regional association rule mining and scoping in spatial datasets

Wei Ding · Christoph F. Eick · Xiaojing Yuan · Jing Wang · Jean-Philippe Nicot

Received: 16 July 2008 / Revised: 30 November 2009 /
Accepted: 31 May 2010 / Published online: 18 June 2010

Abstract The motivation for regional association rule mining and scoping is driven by the facts that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional, rather than global. Furthermore, when using traditional association rule mining, regional patterns frequently fail to be discovered due to insufficient global confidence and/or support. In this paper, we systematically study this problem and address the unique challenges of regional association mining and scoping: (1) region discovery: how to identify interesting regions from which novel and useful regional association rules can be extracted; (2) regional association rule scoping: how to determine the scope of regional association rules. We investigate the duality between regional association rules and regions

Preliminary versions of the paper appeared in [10, 11].

W. Ding (✉)
Department of Computer Science, University of Massachusetts-Boston,
Boston, MA 02125-3393, USA
e-mail: ding@cs.umb.edu

C. F. Eick · J. Wang
Department of Computer Science, University of Houston, Houston, TX 77004, USA

C. F. Eick
e-mail: ceick@uh.edu

J. Wang
e-mail: jwang29@uh.edu

X. Yuan
Engineering Technology Department, University of Houston, Houston, TX 77004, USA
e-mail: xyuan@uh.edu

J.-P. Nicot
Bureau of Economic Geology, John A. & Katherine G. Jackson School of Geosciences,
The University of Texas at Austin, Austin, TX, USA
e-mail: jp.nicot@beg.utexas.edu

where the associations are valid: interesting regions are identified to seek novel regional patterns, and a regional pattern has a scope of a set of regions in which the pattern is valid. In particular, we present a reward-based region discovery framework that employs a divisive grid-based supervised clustering for region discovery. We evaluate our approach in a real-world case study to identify spatial risk patterns from arsenic in the Texas water supply. Our experimental results confirm and validate research results in the study of arsenic contamination, and our work leads to the discovery of novel findings to be further explored by domain scientists.

Keywords Association rule mining and scoping · Region discovery · Clustering · Spatial data mining

1 Introduction

Rapid advances in databases and data acquisition technologies have resulted in an immense amount of spatial data. Efficient knowledge discovery requires various spatial data mining techniques to automatically find novel and useful patterns from large-scale spatial datasets [15, 20, 23, 29, 35, 37, 40, 41, 47]. Of particular interests to scientists is to find scientifically meaningful regions and their associated patterns, for example, identification of earthquake hot spots, revealing high-risk zones that particular cancers associated with environmental pollutions, and the detection of crime zones.

The motivation for regional association rule mining and scoping is driven by the facts that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional, rather than global. It has been pointed out in the literature [19, 33, 39] 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 [19, 31, 32].

Here is an example to illustrate the discrepancies between regional and global associations. Table 1 describes the well data of a county that includes Zone A and Zone B. let us consider an association rule that suggests a well X, up to 251.5-ft deep, is associated with dangerous arsenic concentration as follows,

$$depth(X, 0 - 251.5) \rightarrow arsenic\_level(X, dangerous).$$

Assuming the minimum confidence threshold is 70%, this pattern would not have enough confidence ($\frac{1{,}000}{2{,}000} = 50\% < 70\%$ threshold) to be identified globally in the county. However, the same rule holds in Zone A because its confidence, $\frac{400}{500} = 80\%$, is above the 70% threshold. Notice that this rule does not hold in Zone B, due to its low confidence ($\frac{600}{1{,}500} = 40\%$). Hence a well up to 251.5-ft deep is *positively* associated with high arsenic contamination in Zone A, but is *negatively* associated with dangerous arsenic concentration at the county level. This reversal of an association in the global dataset is also known as spatial heterogeneity [41] or Simpson's Paradox in statistics [14].

Table 1 Contingency tables between well depth and arsenic concentration

	Well depth	Dangerous	Safe	Total
	(0, 251.5]	1,000	1,000	2,000
	(251.5, ∞)	1,200	800	2,000
	Total	2,200	1,800	4,000
ZoneA	(0, 251.5]	400	100	500
	(251.5, ∞)	1,050	450	1,500
ZoneB	(0, 251.5]	600	900	1,500
	(251.5, ∞)	150	350	500
	Total	2,200	1,800	4,000

Unfortunately, traditional association rule mining frequently fails to discover regional patterns due to insufficient global confidence and/or support. A common approach to alleviate the problem is to use a small support threshold. However, this approach usually suffers from a combinatorial explosion in the number of rules generated. Furthermore, for a given dataset, the number of regions as well as the regions themselves are not known *a priori*. This raises two questions: how to measure the interestingness of a set of regions and how to search for interesting regions. One popular approach is to select regions to be mined based on a previously given structure, such as a grid structure using longitude and latitude, or based on political/demographical boundaries, such as counties within a state. But the boundaries of the so-constructed regions usually do not match the natural boundaries of the interesting patterns.

Another unique phenomenon is that regional association rules, by definition, only hold in a subspace but not in the global space; therefore, regional association rules may only be discovered in a particular subspace of the global space. In this paper, we systematically study this problem and address the special challenges for regional association mining and scoping: (1) region discovery: how to identify interesting regions from which novel and useful regional association rules can be extracted; (2) regional association rule scoping: how to determine the scope of regional association rules. Our preliminary work on regional association rule mining was published in [10] and on regional association rule scoping was published in [11]. In this paper, we integrate two originally separated procedures and investigate the duality between regional association rules and regions in which the associations are valid. Interesting regions are identified to seek novel regional patterns, and a regional pattern has a scope that is the set of regions in which the pattern is valid. We design and implement a reward-based framework, utilizing plug-in fitness functions to accomplish two complementary objectives: seeking regions to discover regional association rules, and then identifying regions in which regional association rules are valid. Such regions provide a quantitative measure of how significant a regional association rule is in the global space.

Our contributions To our best knowledge, this paper is the first attempt to propose a structured framework for regional association rule mining and scoping. Firstly, we present a reward-based region discovery framework to search for regions that maximize an external fitness function. We formulate region discovery as a clustering problem to maximize a fitness function that incorporates what domain experts consider interesting. Each cluster is assigned a "reward" value that reflects the cluster's interestingness. In this work, the region discovery framework is used to

identify regions from which regional association rules are mined, and the framework is used to determine the scope of a regional association rule as well. Secondly, we have designed and implemented a new divisive, grid-based supervised clustering algorithm to identify interesting regions, corresponding to different fitness functions. The clustering algorithm searches for clusters to find interesting regions of arbitrary shapes and scales. Finally, we empirically evaluate the effectiveness of our framework in a real-world case study that centers on identifying spatial risk patterns related with arsenic pollution in the Texas water supply. Our experimental results not only confirm and validate research results in the geosciences, but also lead to the discovery of novel findings that need to be further studied by domain scientists.

Figure 1 illustrates the procedure of our approach with a real example from our case study. Interesting regions are identified using a grid-based supervised clustering algorithm and a fitness function designed for the identification of arsenic hot spots. An interesting association rule *a*, *Wells with nitrate concentration lower than 0.085 mg/l have dangerous arsenic concentration*, is discovered from an arsenic hot spot area in the South Texas with 100% confidence. The scope of the association rule *a* is further identified using another fitness function designed for regional association rule scoping. The scope of the associate rule is a larger area that aligns with the Texas Gulf Coast. Further study shows that this regional association rule *a* cannot be discovered at the Texas state level due to its insufficient confidence (less than 50%) on a global scale.

The reminder of this paper is organized as follows. Section 2 reviews related work. Section 3 introduces the framework of regional association rule mining and scoping.

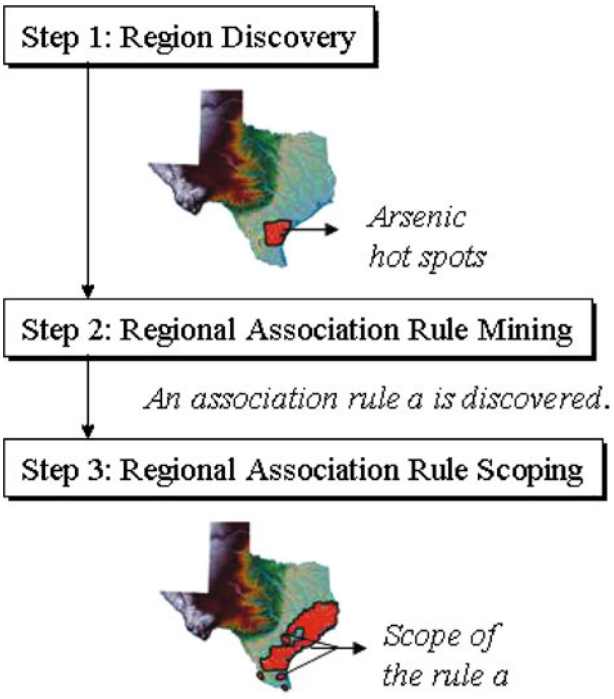


Fig. 1 An example for regional association rule mining and scoping

Section 4 describes the algorithms used in the framework. Section 5 presents the experimental results of a real-world application on identifying arsenic spatial risk patterns in the Texas water supply, and we conclude the paper in Section 6.

2 Related work

The areas most relevant to our work are on hot-spot discovery and spatial association rule mining.

2.1 Hot-spot discovery

Hot spots are traditionally defined as the clusters of "more than usual interest, activity, or popularity" with respect to spatial coordinates [28]. Hot-spot discovery has been investigated in spatial statistics and data mining research.

In spatial statistics, detection of hot spots using a variable resolution approach [7] is investigated to minimize the effects of spatial superposition. In [44], a region-growing method for hot-spot discovery is described, which selects seed points first and then grows clusters from these seed points by adding neighbor points as long as a density threshold is satisfied. The definition of hot spots is extended in [24] using circular zones for multiple variables. Getis and Ord propose a popular method to find hot spots in spatial datasets relying on the G^* Statistic [18, 34]. G^* Statistic detects local pockets of spatial association, and the value of G^* depends on an *a priori* given scale of the packets and is calculated for each object individually. Visualizing the results of G^* calculations graphically reveals hot spots (aggregates of objects with values of G^* higher than expected) and cold spots (aggregates of objects with values of G^* lower than expected). Note that such aggregates are not formally defined clusters since the G^*-based method has no built-in clustering capabilities. Instead, hot spots are inferred from visualization and manual selection.

An alternative approach for hot-spot discovery relies on clustering in data mining. Wang et al. [47] introduce a "region-oriented" clustering algorithm to select hot spots to satisfy certain conditions such as density. Their approach uses statistical information, for example, means and standard deviations, instead of a fitness function to evaluate a cluster. Eick et al. [15, 16] propose Supervised Clustering to maximize cluster purity while keeping the number of clusters low. This paper applies Supervised Clustering to a new problem to find interesting regions (hot spots) that maximize a given fitness function. In this paper, we define two plug-in fitness functions for hot-spot discovery with respect to a class attribute and for identifying the scope of a regional association rule, respectively.

2.2 Spatial association rule mining

Spatial association rule mining [5, 23, 27] applies association rule mining [1] to spatial datasets. Extended from the definition of traditional association rule mining, a spatial association rule takes the form of

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

It denotes an association relation among a set of predicates P_i $(i = 1, ..., m)$ and Q_j $(j = 1, ..., n)$, containing at least one spatial predicate. Spatial predicates may represent topological relations among spatial objects (e.g., intersecting, containing), 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 satisfy both the antecedent and 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 the minimum support and minimum confidence thresholds.

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

1. **Item representation and transaction definition** define "items" and "transactions" for 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 [1] association mining algorithms are often used in Subtasks 2 and 3. These type of algorithms require objects to be described by categorical attributes. Therefore, continuous attributes have to be discretized in Subtask 1, the step of data preprocessing. A transaction is not naturally defined 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 discussed in [23]. Our work adopts the same transaction model.

A daunting problem of spatial association rule mining, especially in real-world applications, is the huge number of generated patterns. Many associations are either already known geographic dependencies or explicitly represented in geographic databases. For example, that gas stations usually locate at road intersections is a well-known and uninteresting association. In order to extract nontrivial and interesting patterns, Borgorny and Sharma et al. [2, 4–6, 38] proposed a set of algorithms to discard previously known and uninteresting associations, using domain knowledge. In particular, the geographic dependencies between the target feature type and a relevant feature type are eliminated to reduce the input space for the frequent itemset generation and previously known and non-interesting geographic dependencies are further removed at the step of frequent itemset generation. To reduce the number of uninteresting patterns, we introduce the concept of Supervised Association Rules (Section 3.1, Definition 1) and seek associations containing the target feature type.

3 The framework for regional association rule mining and scoping

The framework of regional association rule mining and scoping consists of three steps:

Step 1 **Region Discovery:** identifying interesting regions for regional association rules.

Step 2 **Regional Association Rule Mining:** mining regional association rules among discovered regions.

Step 3 **Regional Association Rule Scoping:** determining the scope of regional association rules.

In the remaining part of the section, we will first discuss our reward-based method for region discovery which is closely involved with Steps 1 and 3, and we will formally define the goal of our framework and formulate the measure of interestingness.

3.1 Region discovery

Our region discovery method employs a reward-based evaluation schema that evaluates the quality of the discovered regions. Given a set of regions $R = \{r_1, \ldots, r_k\}$, identified from a spatial dataset $O = \{o_1, \ldots, o_n\}$, the fitness of R, $q(R)$, is defined as the sum of the rewards obtained from each region r_j ($j = 1 \ldots k$):

$$q(R) = \sum_{j=1}^{k} (i(r_j) \times size(r_j)^{\beta}) \tag{1}$$

where $i(r_j)$ is the interestingness measure of a region r_j, a quantity based on domain interest to reflect the degree to which the region is newsworthy. Our reward-based method seeks a set of regions R such that the sum of rewards over all of its constituent regions is maximized. $size(r_j)^{\beta}$ ($\beta > 1$) in $q(R)$ increases the value of the fitness nonlinearly with respect to the number of objects in O belonging to the region r_j. A region reward is proportional to its interestingness, but given two regions with the same value of interestingness, a larger region receives a higher reward to reflect a preference given to larger regions.

We employ clustering algorithms for region discovery. A region is a contiguous subspace that contains a set of spatial objects such that for each pair of objects belonging to the same region, there always exists a path within this region that connects them. We search for regions $r_1, \ldots, r_k$ such that:

1. $r_i \cap r_j = \emptyset$, $i \neq j$, that is, the regions are disjoint.
2. $R = \{r_1, \ldots, r_k\}$ maximizes $q(R)$.
3. $r_1 \cup \ldots \cup r_k \subseteq O$. The generated regions are not required to be exhaustive with respect to the spatial dataset O. It is possible that some objects do not belong to any identified regions; these objects are discarded as outliers due to the lack of interestingness.
4. $r_1, \ldots, r_k$ are ranked based on their reward values. The higher rewards a region receives, the more interesting the region is, with respect to the fitness function q.

3.2 Problem formulation

Let O be a spatial dataset, $S = \{s_1, s_2, ..., s_l\}$ be a set of spatial attributes, $A = \{a_1, a_2, ..., a_m\}$ a set of non-spatial attributes, and $CL = \{cl_1, cl_2, ..., cl_n\}$ 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 the items in O, and 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 $l + m + n$ items). An item $i \in I$ is a binary variable whose value is 1 if the item is presented in t_i ($i = 1, ..., N$) or 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.

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

The rule a holds in the O with the confidence $conf$ and the support sup:

$$sup(P \rightarrow Q) = \frac{|P \cup Q|}{N}$$

$$conf(P \rightarrow Q) = \frac{|P \cup Q|}{|P|}$$

where $|\ |$ denotes the number of elements in a set. A supervised association rule is *strong* if it satisfies user-specified minimum support (min_sup) and minimum confidence (min_conf) thresholds: $sup(P \rightarrow Q) \geq min\_sup$ and $conf(P \rightarrow Q) \geq min\_conf$.

The goal of regional association rule scoping is to compute a set of regions where a given association rule is valid. The scope of a regional association rule represents the spatial impact of this regional pattern. We give formal definition of the scope of an association rule below.

Definition 2 (Scope of an Association Rule) The *scope* of an association rule a is a set of regions in which the association rule a satisfies the min_sup and min_conf thresholds.

Given these definitions and nomenclature, the problem of regional association rule mining and scoping can be formulated as follows.

Find interesting regions, supervised association rules from the discovered regions, and the scope of regional association rules.

Given an itemset I, a classified transaction set T, a set of fitness functions for different measure of interestingness.

3.3 Measure of interestingness

The reward-based framework is designed to support many plug-in interestingness functions, corresponding to various domain interests. The framework utilizes the duality between regions and regional association rules. The framework first identifies "hot" regions using the interestingness function $i_{hotpot\_coldspot}$. After strong regional association rules are identified, the scope of those rules are then calculated, using another interestingness function i_{scope}. Although the same clustering algorithm and the same dataset are used in two different steps, different sets of regions are returned

in two steps due to the different measure of interestingness defined in the fitness functions.

In function $i_{hotpot\_coldspot}$, the measure of interestingness is based on a set of class labels CL. It rewards regions whose probability distribution of CL significantly deviates from its priori probability. A region is a *hot spot*/*cold spot* if its probability distribution of CL is significantly higher / lower than an expected probability. The interestingness function $i_{hotpot\_coldspot}$ is calculated based on $P(r, CL)$ and $priori(CL)$, with the following parameters: η, γ_1, γ_2, R_+, $R\_$, where $\eta > 0, \gamma_1 \leq 1 \leq \gamma_2$, $0 \leq R_+, R\_ \leq 1$. $P(r, CL)$ is the probability of objects in a region r belonging to CL; $priori(CL)$ is the probability of objects in the global dataset O belonging to CL; R_+ and $R\_$ are the maximum rewards for hot spots and cold spots, respectively.

$$i_{hotpot\_coldspot} = \begin{cases} [\dfrac{priori(CL) \times \gamma_1 - P(r, CL)}{priori(CL) \times \gamma_1} \times R\_]^{\eta} & if\ P(r, CL) < priori(CL) \times \gamma_1 \\ [\dfrac{P(r, CL) - priori(CL) \times \gamma_2}{1 - priori(CL) \times \gamma_2} \times R_+]^{\eta} & if\ P(r, CL) > priori(CL) \times \gamma_2 \\ 0 & otherwise \end{cases} \quad (2)$$

The parameter η determines how quickly the value of interestingness grows to the maximum value (either R_+ or $R\_$). If η is set to 1, the interestingness function changes linearly, as shown in Fig. 2. In general, the larger the value for η is, the higher rewards for purer clusters are. $priori(CL) \times \gamma_1$ and $priori(CL) \times \gamma_2$ determine the thresholds based on which a reward is given to a cluster.

The following example explains how to calculate the fitness of a clustering schema X of an example dataset using Eqs. 1 and 2.

Example Let us assume a clustering schema R is evaluated with respect to the class of interest *dangerous* (high-level arsenic) concentration with $priori(dangerous) = 0.2$ and a dataset that contains 1,000 examples. Suppose that the dataset is partitioned into four clusters denoted as $X = \{x_{11}, x_{12}, x_{13}, x_{14}\}$, and $|x_{11}| = 50$, $|x_{12}| = 200$, $|x_{13}| = 400$, $|x_{14}| = 350$. Assume that there are 20, 100, 80, and 0 objects labeled "dangerous" in the four 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: $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 1$. Hence, $priori(CL) \times \gamma_1 = 0.2 \times 0.5 = 0.1$, and $priori(CL) \times \gamma_2 = 0.2 \times 1.5 = 0.3$. With this setting, a cluster does not receive any reward if its

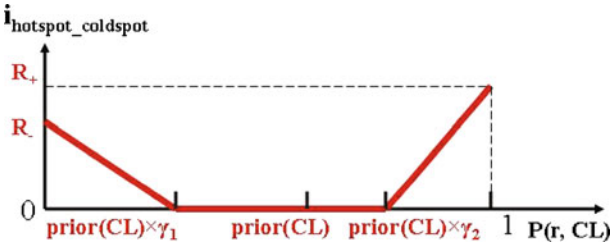


Fig. 2 The interestingness function $i_{hotpot\_coldspot}$ using $\eta = 1$

probability of class "dangerous" is not significantly higher or lower than the expected probability, that is, the value is between $priori(CL) \times \gamma_1 = 0.1$ and $priori(CL) \times \gamma_2 = 0.3$. Therefore, x_{13} receives no reward. The interestingness for the other clusters using $\eta = 1$ is

$$i_{hotpot\_coldspot}(x_{11}) = \left(\frac{0.4 - 0.3}{1 - 0.3}\right)^1 = \frac{1}{7},$$

$$i_{hotpot\_coldspot}(x_{12}) = \left(\frac{0.5 - 0.3}{1 - 0.3}\right)^1 = \frac{2}{7},$$

$$i_{hotpot\_coldspot}(x_{14}) = \left(\frac{0.1 - 0}{0.1}\right)^1 = 1.$$

The fitness value of the clustering schema X calculated using Eq. 1 with $\beta = 1.1$ is

$$\begin{aligned} q(X) &= \frac{1}{7} \times \left(\frac{50}{1{,}000}\right)^{1.1} + \frac{2}{7} \times \left(\frac{200}{1{,}000}\right)^{1.1} \\ &\quad + 0 \times \left(\frac{400}{1{,}000}\right)^{1.1} + 1 \times \left(\frac{350}{1{,}000}\right)^{1.1} \\ &= 0.369 \end{aligned}$$

Function i_{scope} evaluates the interestingness of a region for a given association rule. Let a be an association rule, $conf(a, r)$ the confidence of a in a region r, and $sup(a, r)$ the support of a in r, we define the interestingness $i_{scope}(r)$ of region r with respect to the given association rule a as follows:

$$i_{scope}(r) = \begin{cases} 0 & \text{if } sup(a, r) < min\_sup \times \delta_1 \text{ or} \\ & conf(a, r) < min\_conf \times \delta_2, \\ \left(\dfrac{sup(a, r)}{min\_sup}\right)^{\eta_1} \left(\frac{conf(a,r) - min\_conf \times \delta_2}{1 - min\_conf \times \delta_2}\right)^{\eta_2} & \text{otherwise.} \end{cases} \tag{3}$$

In regional association rule scoping, a region's reward is proportional to its interestingness, which is determined based on the confidence and support of association rule a in region r. In Eq. 3, the thresholds $min\_sup \times \delta_1$ and $min\_conf \times \delta_2$ are introduced to weed out regions in which the association a barely holds. The minimum support and confidence thresholds prevent the clustering solution from containing large clusters with low interestingness. Values of parameters η_1 and η_2 ($\eta_1, \eta_2 > 0$) determine the weight to the increment of the support and confidence, respectively.

The measure of interestingness defined in i_{scope} uses "soft" instead of "hard" thresholds to avoid a harsh crisp effect [3]. For example, with $\delta_1 = \delta_2 = 0.9$, the function $i_{scope}(r)$ rewards regions as long as their confidence or support thresholds are within 90% of the hard thresholds min_conf and min_sup. For example, let's assume that $min\_sup = 10\%$, $min\_conf = 80\%$, and that the association rule under consideration has support = 9% and confidence = 100% in a region r'. In this case, instead of assigning zero reward to region r', we argue to reward the region because the confidence of the rule in region r' is significantly above the min_conf threshold and its support is just a little bit lower (1%) than the min_sup threshold. Our approach uses a quantitative evaluation method that assigns a higher degree of interestingness

and consequently a higher reward to regions whose support and confidence are high with respect to an association rule of interest. Once an association rule a is discovered from a particular region r in the first place, we know that region r from which the association rule a originates, receives a positive reward due to the fact that a satisfies the support and confidence thresholds in r. Consequently, region r will always be contained in the set of regions that define the scope of association rule a.

4 Algorithms

4.1 Region discovery

We formulate region discovery as a clustering problem to search for clusters that maximize domain-specific metrics as described in detail in previous section. Different measure of interestingness may lead to different sets of identified regions. Consequently, clustering algorithms embedded in the framework should allow for plug-in fitness functions. However, the use of fitness functions is quite uncommon in clustering methods, although a few exceptions exist, for example, the hierarchical clustering algorithm CHAMELEON [22] uses fitness functions to evaluate interconnectivity and proximity between two clusters. Furthermore, our region discovery method is different from traditional clustering methods as it is geared toward finding interesting places with respect to a given measure of interestingness. Clusters are ranked based on reward values, and clusters receive low rewards are discarded as outliers and will not be identified as interesting regions.

We have designed and implemented a new Supervised Clustering algorithm using Multi-Resolution Grids (SCMRG). SCMRG is a hierarchical, grid-based method that utilizes a top-down search. The spatial space of the dataset is partitioned into grid cells. Each grid cell at a higher level is partitioned further into smaller cells at the lower level, and this process continues as long as the sum of the rewards of the lower level cells $q(R)$ is not decreased. The regions returned by SCMRG are the combination of grid cells obtained at different levels of resolution. The number of clusters, k, is calculated by the algorithm itself.

Algorithm 1 gives the pseudo-code of SCMRG. A queue data structure is used to store all the cells that need to be processed. The algorithm starts at a user-defined level of resolution and considers the following three cases when processing a cell c:

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 also greater than the sum of rewards of its grandchildren, this cell is returned as a cluster by the algorithm (steps 15–17).

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 descendants will be labeled clusters (steps 23–29).

Case 3 otherwise, put all the children of the cell c ($succ(c)$) into a queue for further processing (steps 18–21, steps 24–28).

The algorithm traverses through the hierarchical structure and examines those cells in the queue from the higher level. It uses a user-defined cell size as a depth boundary. Cells smaller than this cell size will not be split any further (step 19, step

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

Input:

- A fitness function.
- A level of resolution l for the initial grid structure.
- The minimum cell size. A cell will not be divided further if it approaches the minimum cell size.

Output:

- Discovered regions $R = \{r_1, \ldots, r_k\}$.

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 the 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 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.

25). Finally, SCMRG collects all the cells that have been identified in Case 1 from different levels, and merges neighbor clusters if the overall fitness can be improved. The obtained regions are returned as the result of the SCMRG clustering algorithm (steps 31–33).

This hierarchical grid-based approach captures clustering information associated with spatial cells without recourse to the individual objects because we do not drill down a cell if it does not look so promising (Case 2). SCMRG avoids time-consuming distance calculation because it uses the grid structure to define the neighborhood of objects. The computational complexity of SCMRG is thus linear in 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 SCMRG algorithm has some similarity with the STING clustering algorithm [47]. The difference is that the SCMRG algorithm focuses on finding interesting cells (those receive high rewards) instead of cells that contain answers to a given query. In addition, SCMRG only computes cell statistics when needed and not in advance as STING does, thus saving storage space as well.

The complexity of the SCMRG algorithm is controlled by two factors: the number of the candidate cells in the queue and the calculation of the fitness. The algorithm calculates the fitness of all objects inside a cell and a cell will not be further divided if drilling down cannot improve the current reward. The number of cells of a layer is less than one-forth of the number of the layer one level lower. The total number of cells to be processed in the worst case is less than $1.33N_c$, where N_c is the number of the cells at the bottom layer.[1] The actual number of cells is usually less than $1.33N_c$ due to the reward-based pruning. It is also reasonable to assume that each cell at the bottom layer likely contains many objects because the reward function is designed to favor larger cell (a.k.a. larger clusters). In our empirical study, the average cell size is above 400 objects. In general, the total number of cells is much less than the total number of objects. Let the cost of fitness calculation is $O(q)$. Thus the complexity of the algorithm in average is usually much better than $O(N) \times O(q)$, where N is the total number of objects in the dataset.

The example in Fig. 3 explains the procedure of the SCMRG algorithm using a sample dataset. The first decomposition results into four cells c_{11}, c_{12}, c_{13}, c_{14} at Level 1. If the reward of c_{11} is greater than the sum of the rewards of its children, and if it is also greater than the sum of rewards of its grandchildren, c_{11} is then labeled a cluster according to Case 1. Cell c_{14} does not receive any rewards, if neither its children nor grandchildren receive any rewards. According to Case 2, c_{14} is not labeled a cluster, and its successors are not saved in the queue. Although Cell c_{13} receives no reward, assume its children receive rewards, all the children of c_{13} are saved in the queue to be further processed (Case 3). The cells at Level 1 are then divided into Levels 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 Levels 2 and 3 in Fig. 3. Neighbor clusters are merged if this improves the fitness. In this example, two regions are identified.

[1] Total number of cells $= N_c \times (1 + \sum_{n->\infty} \frac{1}{4^n})$ and $\sum_{n->\infty} \frac{1}{4^n} = \frac{1}{3}$.

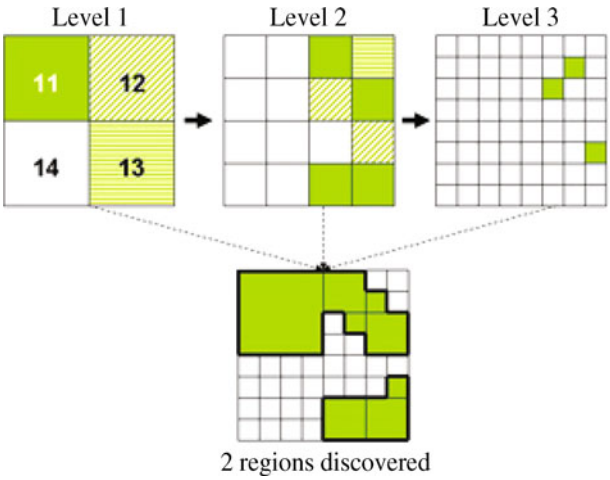


Fig. 3 Running the SCMRG algorithm on a sample dataset

4.2 Generation of regional association rules

Once regions are identified, we construct frequent itemsets for each region. Our Supervised_Apriori_Gen algorithm (see Algorithm 2) extends the *Apriori* algorithm [1] by utilizing a given class structure.

The *Apriori* algorithm first makes a single pass over the dataset 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. A k-itemset is an itemset that has k attributes. A candidate itemset is pruned if it is not frequent. The algorithm terminates when there are no new frequent itemsets generated, for example, $F_k = \emptyset$. Supervised_Apriori_Gen algorithm uses a different approach: the given class structure is incorporated by enforcing that each candidate k-itemset include at least one class label; otherwise it is pruned even if it is frequent. The Supervised-Apriori-Gen uses the $F_{k-1} \times F_{k-1}$ method [43] to merge a pair of frequent (k-1)-itemsets. 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)-itemsets. A and B are merged to form a k-itemset $\{a_1, a_2, ..., a_{k-1}, b_{k-1}\}$ (see *form* function in step 22) if they satisfy the following conditions:

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

The Supervised-Apriori-Gen algorithm initially starts with a candidate 2-itemset construction, which is the basis of the k-itemset generation ($k > 2$). To ensure that each 2-itemset includes at least one class label, the algorithm firstly constructs candidate 1-itemsets from frequent 1-itemsets (steps 2–4). The algorithm separates class-label items from other items using the *split* function (step 5). Next the algorithm enumerates class-label items with the rest of items (steps 6–11), as well as class-label items with themselves (steps 12–18). Thus, steps 6–11 generate candidate 2-itemsets formed between class labels and other non-class-label items; steps 12–17 generate candidate 2-itemsets formed between class labels. The 2-itemsets are then used for k-itemsets generation ($k > 2$) (steps 19–26).

Algorithm 2 Supervised_Apriori_Gen: Candidate Generation and Pruning

Input:

- k-1 frequent itemset C_{k-1}.
- The minimum support threshold.

Output:

- k candidate itemset C_k.

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. $(C_{1\_class\_label}, C_{1\_other}) = split(C_1, CL)$.
6. **for** each candidate itemset $c1 \in C_{1\_class\_label}$ **do**
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 $c1 \in C_{1\_class\_label}$ **do**
13. **for** each candidate itemset $c2 \in C_{1\_class\_label} - \{c1\}$ **do**
14. c = form $c1$ and $c2$.
15. insert c into C_2.
16. **end for**
17. **end for**
18. **else**
19. **for** each $i1$ in F_{k-1}
20. **for** each $i2$ in F_{k-1}
21. **if** (first $k-2$ items of $i1, i2$ are same) $\wedge$ (last item of $i1, i2$ differs)
22. c = form (first $k-1$ items of $i1$) and (last item of $i2$).
23. insert c into C_k.
24. **end if**
25. **end for**
26. **end for**
27. **end if**
28. return C_k.

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

5 Arsenic regional association rule mining and scoping in the Texas water supply

In this section, we describe the experimental procedures of applying the framework of regional association rule mining and scoping to a real world case study that

identifies arsenic spatial risk patterns in the Texas water supply. We then 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 cleaning data, transforming continuous attributes into categorical attributes, and constructing transactions using water wells as the reference feature.
2. Identifying arsenic *hot spots* and *cold spots*. A region whose arsenic distribution is significantly higher than the Texas state level is considered an arsenic hot spot; a region whose arsenic distribution is significantly lower the Texas state level is considered an arsenic cold spot.
3. Mining supervised association rules from each identified region and for the complete dataset.
4. Determining scope of strong supervised association rules.

5.1 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 [45]. The Texas Water Development Board has monitored and analyzed arsenic concentration over the last 30 years. Arsenic in very high concentration is poisonous. Long term exposure to arsenic, even though at low level, can still lead to increased risk of cancers [42]. Arsenic is derived from both anthropogenic sources, such as the drainage from mines and mine tailings, pesticides, and biocides, and from natural sources, such as the 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 [48], as one of the key parameters for drinking water quality and safety evaluation.

Because data collection and maintenance procedures and standards have changed over the years in 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 nine-intersection model [13]. Non-spatial attributes are selected with the assistance of domain experts [21, 25, 36]; they include *well depth*, and concentration of *fluoride*, *nitrate*, and other chemical metal elements including *vanadium*, *iron*, *molybdenum* and *selenium*. Among those attributes, the attribute *well depth* is used for studies on mobilizing mechanism; the attributes *vanadium* and *molybdenum* have similar geochemical behavior; the attributes *fluoride*, *nitrate*, *iron*, and *selenium* may suggest the ultimate origin of arsenic. The arsenic dataset generated by our research group and the dataset is available on the web at [9].

We classify water wells into two classes: *safe* and *dangerous*. Based on the standard for drinking water defined by the Environment Protection Agency [46], a well is considered dangerous if its arsenic concentration level is above 10 μg/l. To ensure the quality of the association rule generated in the study, we only select lab test results that use honored sampling procedures. This results in 11,922 records selected from

GWDB after data preprocessing. Figure 4 illustrates arsenic contamination in Texas, where dangerous wells are in red (or dark gray).

Table 2 describes the seven non-spatial attributes used in the arsenic dataset. The table lists the mean and the standard deviation of those continuous attributes before discretization. In preparation of the association rule mining, continuous attributes excluding latitude and longitude are first converted into categorical attributes. In general, two different methods are used for discretization of continuous attributes: unsupervised discretization without using class information and supervised discretization using class information [43]. In our experiments, we adopt the supervised method Recursive Minimal Entropy Partitioning introduced in [17]. The supervised entropy-based method uses class labels *dangerous* and *safe* to place the splits in a way that maximizes the purity of arsenic classes in the intervals. This discretization method maximized the support for arsenic class attribute, facilitating the discovery of supervised association rules involving with arsenic. Hence the method can effectively find the supervised association rules related with arsenic classes. The method produces unequal bin sizes and has been shown to produce better results in data mining tasks [12]. The splitting points of each continuous attribute are listed in Table 2. For example, the value of nitrate concentration has been discretized into five intervals with respect to the arsenic classes: (0,0.085], (0.085,0.455], (0.455,16.1], (16.1,28.085], and (28.085,∞) (measurement unit mg/l).

5.2 Region discovery for arsenic hot/cold spots

We have re-discovered several interesting risk regions with high arsenic concentration (hot spots), which have been studied by geoscientists before. We have also identified regions with low arsenic concentration (cold spots). The association rules that we constructed from those identified regions can help geoscientists identify the causes of high arsenic concentration in different regions. We now present our

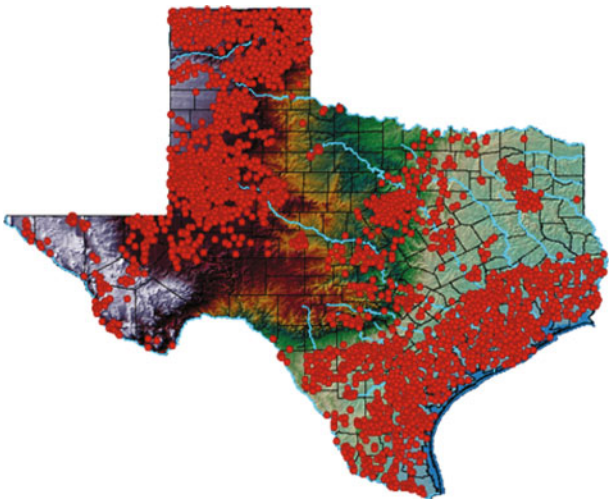


Fig. 4 Arsenic contamination in Texas; background depicts Texas terrain color ramp. Legend: *red (or dark gray) dots*—dangerous wells

Table 2 Arsenic dataset

Non-spatial attributes	Mean	STD	Splitting points
1. well depth (foot)	587.959	654.962	215.5
2. nitrate (mg/l)	11.362	27.499	0.085, 0.455, 16.1, 28.085
3. fluoride (mg/l)	1.161	1.349	0.315, 2.445, 3.375, 4.605
4. vanadium ($\mu g/l$)	8.755	25.827	1.2, 2.05, 2.95, 3.25, 5.945, 11.85, 19.95, 20.05, 37.95
5. iron ($\mu g/l$)	9.226	15.651	1.295, 2.595, 4.945, 5.015, 7.895, 19.65, 20.05, 48.05, 51.75
6. molybdenum ($\mu g/l$)	259.882	1,320.784	9.05, 11.35, 19.95, 20.1, 28.1, 47.2, 51.05
7. selenium ($\mu g/l$)	14.243	34.75	4.995, 5.01, 19.95, 20.05, 34.65, 43.05, 52.85, 74.55
Total # of wells	11,922		

results with validation from the published results in the geosciences for both region discovery and association rule mining and scoping.

In region discovery, the SCMRG algorithm is applied to the dataset that consists of longitude and latitude of wells along with arsenic class labels *dangerous* or *safe* using Eq. 2. Figure 5 depicts the result of the top four regions that have received the highest reward. Specifically, Regions 1 and 3 have high density of dangerous wells, and Regions 2 and 4 have high density of safe wells. Hot spot Region 1 overlaps with the arsenic risk zone reported in the National Water-Quality Assessment Program [30], and hot spot Region 3 is confirmed as an arsenic risk zone by Parker's work [36].

If we are interested in finding larger regions with lower purity, using a larger value of β results in a bigger size of the 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 $\beta = 1.01$ and $\eta = 1$ give us the best results in the rules constructed in supervised association rule mining.

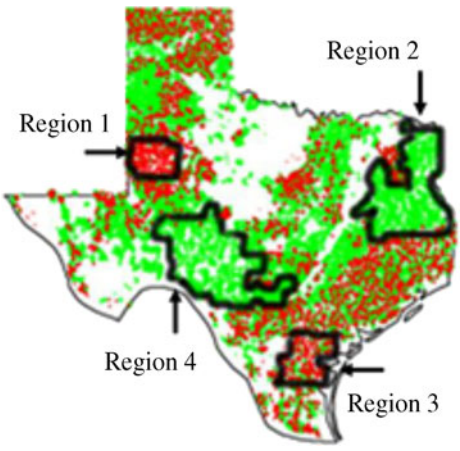


Fig. 5 Interesting regions are identified using $\beta = 1.01$, $\eta = 1$, $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 1$. Average region purity = 0.85

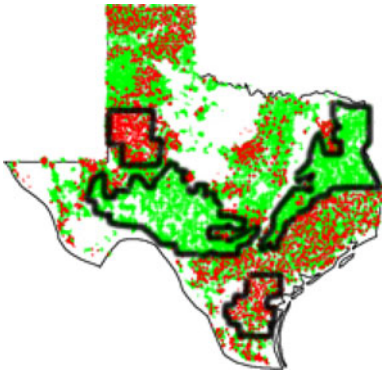


Fig. 6 Interesting regions are identified using $\beta = 1.035$, $\eta = 1$, $\gamma_1 = 0.5$, $\gamma_2 = 1.5$, $R+ = 1$, $R- = 1$. Average region purity = 0.83

5.3 Regional 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 $min\_confidence = 70\%$ thresholds for the experiments. We present the first few rules for the regions investigated, which are all meaningful and important according to the arsenic study literature.

Mining regional rules in arsenic hot spots discovers attributes that are associated with high arsenic concentration; in cold spots it discovers attributes related to low arsenic concentration. For example, in Region 3 of Fig. 5, we discover

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

The rule states, with 100% confidence, that the wells in Region 3 with nitrate concentration lower than 0.085 mg/l have dangerous arsenic concentration. The strong association between nitrate and high arsenic concentration is verified by Hudak's work [21] in environmental geology.

In Region 1 of Fig. 5, we also discover

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

The rule states with 100% confidence that the wells in Region 1, with vanadium concentration between 20.05 and 37.95 μg/l and selenium concentration larger than 74.55 μg/l, have dangerous arsenic concentration. Our discovery is confirmed by the work of Lee et al. in [25].

Our experimental results also show some novel rules that have not been reported in the literature of arsenic analysis. For example, in Region 1 the following rule is discovered:

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

The rule indicates that shallow wells with a certain range of iron concentration are associated with high arsenic concentration. We hope that the results from our study

will help domain experts in selecting interesting hypotheses for further scientific exploration.

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 (hot spots), as well as for Region 2 and Region 4 (cold spots). Due to different geographical structure and farm activities of the study area, the spatial risk patterns associated with arsenic are different in each region. For example, comparing the previously studied rule 1 identified in Region 3 with rule 4 extracted from Region 1:

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

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

Rules in Regions 2 and 4 (cold spots) shed light on what may prevent high arsenic concentration. For example, we find the following rule, discovered both in Regions 2 and 4, states what is associated with low arsenic concentration.

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

For comparative purposes, we also mine supervised association rules in the whole dataset. Using low support values in global datasets to find more interesting association rules has been suggested by [26]. However, even with a rather low support threshold $min\_support = 1\%$, none of the top ranked interesting regional association rules we identified previously are included among over 100,000 resulting rules. On the other hand, up to 300 rules on average are identified per region using our framework with $min\_support = 10\%$ and $min\_confidence = 70\%$ thresholds. Regional association rules identified from those arsenic hot/cold spots tend to be more revealing and interesting. Not surprisingly, a large portion of 100,000 statewide association rules are trivial and general rules, such as

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

This global association rule claims that wells which are used by human beings and have safe arsenic concentration are very likely (confidence is 86%) located in river basin 19 (in San Antonio area). It is a well-known fact in Texas.

5.4 Region discovery for regional association rule scoping

We use the same clustering algorithm SCMRG but a different fitness function i_{scope} (Eq. 3) for regional association rule scoping. The following four regional association rules with 100% confidence from Regions 1, 2, 3, and 4 are used as illustration

examples in the rest of this section for regional association rule scoping. Association rules 1 and 3 are confirmed in arsenic literature [21, 25].

Association Rule 1

$$nitrate(X, 28.31 - \infty) \wedge arsenic\_level(X, dangerous) \rightarrow depth(X, 0 - 251.5)$$

Association Rule 2

$$depth(X, 0 - 251.5) \wedge fluoride(X, 0 - 0.085) \rightarrow arsenic\_level(X, safe)$$

Association Rule 3

$$nitrate(X, 0 - 0.085) \rightarrow arsenic\_level(X, dangerous)$$

Association Rule 4

$$depth(X, 251.5 - \infty) \wedge nitrate(X, 0.265 - 16.1) \rightarrow arsenic\_level(X, safe)$$

Figure 7 depicts the scope of four association rules above. The scope of an association rule can contain several regions. The scope of Association Rule 1 (top row, left column) overlaps with the Texas High Plains. In this area, shallow depth wells (<251.5 ft) indicate the aquifer is thin; thus, nitrate comes from surface contamination (>28.31 MG/L). Arsenic contamination is of geological origin and is then enhanced by the lack of dilution because the aquifer is thin. The scope of Association Rule 3 (bottom row, left column) is applicable to the whole Texas Gulf Coast because the geology there is similar. The scope of Association Rules 2 and 4

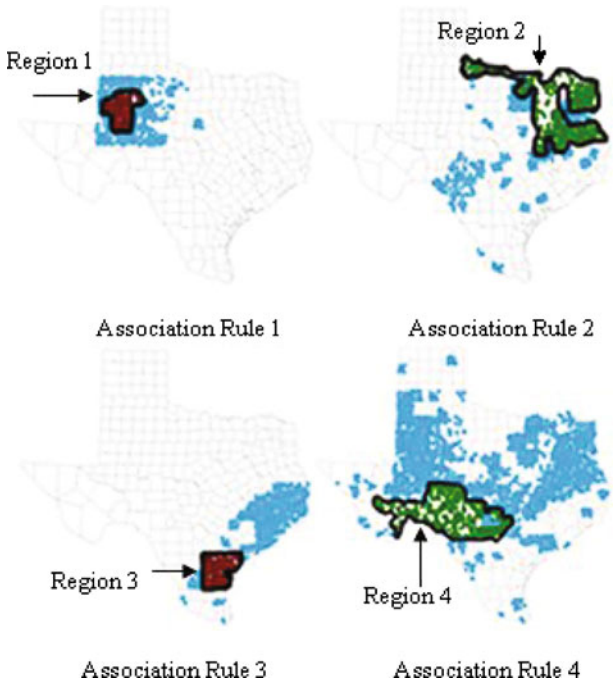


Fig. 7 Region–Regional association rule–Scope using $\beta = 1.01$, $\eta_1 = 1, \eta_2 = 1.1$, $\delta_1 = \delta_2 = 0.9$, $min\_sup = 10\%$, $min\_conf = 80\%$. Legend: regions are highlighted by *bold border line*; scopes are in *color blue* (or *light grey*)

represents the areas where arsenic contamination is low. They are interesting places that domain scientists will explore in the future.

It is also important to point out that the scope of an association rule indicates how global, regional, or local a pattern is. For example, the scope of the association rule 4 in Fig. 7 covers a large percentage of the global space ($> 75\%$). We find that the association rule 4 is also valid (holds with 85% confidence) in the global dataset. Hence, it is indeed a global association rule. However, none of the other three association rules are discovered globally. We can also fine-tune the measure of interestingness for association rule scoping by varying its support and confidence thresholds for a given association rule. Figure 8 shows how the scope of the association rule 3 changes using different confidence and support thresholds. Typically, a lower value of the *min_sup* results in a larger scope; a higher value of the *min_conf* results in a smaller scope.

Association rule scoping has many applications that go beyond the proposed framework introduced in this paper. Scoping can be applied to any spatial association rules, including global association rules. For example, a domain expert can check whether an arsenic association, which is valid in Texas, also holds in Bangladesh, a country that has serious arsenic contamination in drinking water. It is also inspirational for domain experts to explore how the scope of an association rule changes, if an association rule is slightly modified, for example, a condition in its antecedent is dropped. Furthermore, in addition to finding the scope where an association holds, it might be interesting to search for the scope where it does not hold. For example, if we find that high levels of iron associates with high arsenic concentration in one region, but with low arsenic concentration in another region, this case should be further analyzed. Last but not least, the regions obtained using association rule

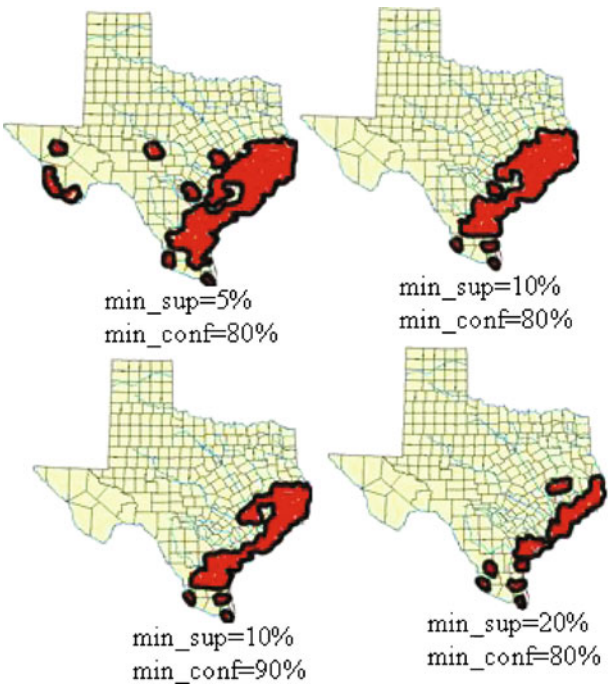


Fig. 8 The scope of a particular rule changes based on the different values of *min_sup* and *min_conf*. $\beta = 1.01,\ \eta_1 = 1, \eta_2 = 1.1,\ \delta_1 = \delta_2 = 0.9,\ min\_sup = 10\%,\ min\_conf = 80\%$

scoping can serve as a source for mining new association rules. For example, if we are interested in the places where high levels of iron associate with high levels of fluoride, $high\_iron(X) \rightarrow high\_fluoride(X)$. We can then determine the scope of this association rule and use the new obtained regions to mine new interesting association rules that provide further details that contribute to the association between iron and fluoride.

Our SCMRG algorithm is computationally efficient. On average, it takes 3.031 s for hot spots/cold spots discovery, and 4.68 s for regional association rule scoping. The computer has an Intel(R) Pentium(R) M, a CPU 1.2 GHz, and 632 MB of RAM. The algorithm implemented in Java can be accessed on the Web at our open source project *Cougar*[2] *Java Library for Machine Learning and Data Mining Algorithms* [8].

6 Discussion

One critical requirement for spatial data mining is the capability to analyze datasets at different levels of granularity, as well as analyze the data globally. We face two unique challenges in regional association mining and scoping: (1) how to determine regions from which regional association rules will be extracted, and (2) how to compute the scope of regional association rules. We solve the first issue using a reward-based region discovery algorithm that employs a grid-based supervised approach to identify interesting subregions in spatial datasets. We address the second problem by exploiting the duality between regional patterns and regions: regions are used to discover regional association rules; next the obtained regional association rules are used to determine places in which the association rules are valid. Such regions capture the scopes of regional patterns and provide a quantitative measure of how significant a regional association rule is in the global space.

We evaluate the proposed framework in a real-world case study to identify spatial risk patterns and risk zones of arsenic in the Texas water supply. The goal of the case study is to understand what regional associations exist between high arsenic concentration and other factors. We have identified arsenic hot spots and cold spots, created regional rules from the obtained regions, and evaluated the spatial impact of interesting regional associations. We are not interested in predicting whether a well is safe or dangerous because this information is already known. A classification algorithm would only be helpful if we could drill into the classification model to determine which factors are associated with high arsenic pollution. In general, our work can be viewed as an exploratory data analysis approach that centers on which features are potentially relevant in causing arsenic pollution. Moreover, our approach identified several new relationships between arsenic and other factors which provide scientists with novel hypotheses for further exploration.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S (eds) Proceedings of the 1993 ACM SIGMOD international conference on management of data, Washington, D.C., vol 26–28, pp 207–216
2. Appice A, Ceci M, Lanza A, Lisi FA, Malerba D (2003) Discovery of spatial association rules in geo-referenced census data: a relational mining approach. Intell Data Anal 7(6):541–566

3. Bistarelli S, Bonchi F (2005) Interestingness is not a dichotomy: introducing softness in constrained pattern mining. In: The ninth European conference on principles and practice of knowledge discovery in databases (PKDD). Lecture notes in computer science, vol 3721. Springer, Porto, Portugal
4. Bogorny V, Camargo S, Engel PM, Alvares LO (2006) Mining frequent geographic patterns with knowledge constraints. In: GIS '06: proceedings of the 14th annual ACM international symposium on advances in geographic information systems, Arlington, Virginia, USA, pp 139–146
5. Bogorny V, Kuijpers B, Alvares L (2008) Reducing uninteresting spatial association rules in geographic databases using background knowledge: a summary of results. Int J Geogr Inf Sci 22(4):361–386
6. Bogorny V, Valiati J, Camargo S, Engel P, Kuijpers B, Alvares L (2006) Mining maximal generalized frequent geographic patterns with knowledge constraints. In: The 6th international conference on data mining, Hong Kong, pp 813–817
7. Brimicombe AJ (2005) Cluster detection in point event data having tendency towards spatially repetitive events. In: The 8th intl. conf. on GeoComputation
8. CougarSquared Data Mining and Machine Learning Framework, Data Mining and Machine Learning Group (2009) University of Houston. http://cougarsquared.dev.java.net/
9. Data Mining and Machine Learning Group (2009) University of Houston. http://www.tlc2.uh.edu/dmmlg/Datasets
10. Ding W, Eick CF, Wang J, Yuan X (2006) A framework for regional association rule mining in spatial datasets. In: The 6th IEEE international conference on data mining (ICDM)
11. Ding W, Eick CF, Yuan X, Wang J, Nicot J-P (2007) On regional association rule scoping. In: The international workshop on spatial and spatio-temporal data mining in cooperation with IEEE ICDM 2007, Omaha, NE, USA
12. Dougherty J, Kohavi R, Sahami M (1995) Supervised and unsupervised discretization of continuous features. In: International conference on machine learning, pp 194–202
13. Egenhofer MJ, Franzosa RD (1991) Pointset topological spatial relations. Int J Geogr Inf Syst 5(2):161–174
14. EH S (1951) The interpretation of interaction in contingency tables. J R Stat Soc B13:238–241
15. Eick C, Vaezian B, Jiang D, Wang J (2006) 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
16. Eick CF, Zeidat N, Zhao Z (2004) Supervised clustering: Algorithms and application. In: International conference on tools with AI, Boca Raton, Florida, pp 774–776
17. Fayyad UM, Irani KB (1993) Multi-interval discretization of continuous-valued attributes for classification learning. In: Kaufmann M (ed) Proceedings of the 13th international joint conference on artificial intelligence, pp 1022–1027
18. Getis A, Ord JK (1992) The analysis of spatial association by use of distance statistics. Geogr Anal 24:189–206
19. Goodchild MF (2003) The fundamental laws of GIScience. Invited talk at University Consortium for Geographic Information Science, University of California, Santa Barbara
20. Han J, Kamber M, Tung AKH (2001) Spatial clustering methods in data mining: a survey. In: Geographic data mining and knowledge discovery
21. Hudak PF (2003) Arsenic, nitrate, chloride and bromide contamination in the gulf coast aquifer, south-central Texas, USA. Int J Environ Stud 60:123–133
22. Karypis G, Han E-HS, Kumar V (1999) Chameleon: hierarchical clustering using dynamic modeling. IEEE Computer 32(8):68–75
23. Koperski K, Han J (1995) Discovery of spatial association rules in geographic information databases. In: Egenhofer MJ, Herring JR (eds) Proc. 4th int. symp. advances in spatial databases, SSD, vol 951, pp 47–66, 6–9
24. Kulldorff M (2001) Prospective time periodic geographical disease surveillance using a scan statistic. J R Stat Soc Ser A 164:61–72
25. Lee LM Herbert B (2001) 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)
26. Li W, Han J, Pei J (2001) CMAR: accurate and efficient classification based on multiple class-association rules. In: International conference on data mining (ICDM'01), San Jose, CA
27. Mennis J, Liu J (2005) Mining association rules in spatio-temporal data: an analysis of urban socioeconomic and sand cover change. Trans GIS 9:5–17

28. Merriam-Webster Online Dictionary (2009) http://www.merriam-webster.com
29. Munro R, Chawla S, Sun P (2003) Complex spatial relationships. In: The third IEEE international conference on data mining (ICDM)
30. National Water-Quality Assessment Program, U.S. Department of the Interior and U.S. Geological Survey (2001) Ground-water quality of the southern high plains aquifer, Texas and New Mexico, Open-File Report 03-345
31. Openshaw S (1994) Two exploratory space–time attribute pattern analysers relevant to GIS. In: Fotheringham S, Rogerson P (eds) Spatial analysis and GIS. Taylor and Francis, London, pp 83–104
32. Openshaw S (1995) Developing automated and smart spatial pattern exploration tools for geographical information systems applications. Statistician 44(1):3–16
33. Openshaw S (1999) Geographical data mining: key design issues. In: GeoComputation
34. Ord JK, Getis A (1995) Local spatial autocorrelation statistics: distributional issues and an application. Geogr Anal 27(4):286–306
35. Papadimitriou S, Gionis A, Tsaparas P, Väisänen A, Mannila H, Faloutsos C (2005) Parameter-free spatial data mining using MDL. In: 5th international conference on data mining (ICDM)
36. Parker R (2001) 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
37. Roddick JF, Spiliopoulou M (1999) A bibliography of temporal, spatial and spatio-temporal data mining research. In: SIGKDD explorations, vol 1, pp 34–38
38. Sharma L, Tiwary U, Vyas O (2004) An efficient approach to spatial association rule mining. In: Int. conf. on ISPR IIIT, Allahabad, India, pp 1–5
39. Shekhar S (2004) Spatial data mining: accomplishments and research needs. In: Keynote speech at GIScience 2004 (3rd bi-annual international conference on geographic information science)
40. Shekhar S, Chawla S (2003) Spatial databases: a tour. Prentice Hall, Upper Saddle River (ISBN 013-017480-7)
41. Shekhar S, Zhang P, Huang Y, Vatsavai RR (2003) Book chapter in data mining: next generation challenges and future directions. In: Kargupta H, Joshi A (eds) Spatial data mining. AAAI/MIT, Cambridge
42. Smith A, Hopenhayn-Rich C (1992) Cancer risks from arsenic in drinking water. In: Environmental health perspectives, vol 97, pp 259–267
43. Tan P-N, Steinbach M, Kumar V (2006) Introduction to data mining. Addison-Wesley, New York
44. Tay SC, Hsu W, Lim KH (2003) Spatial data mining: clustering of hot spots and pattern recognition. In: IEEE international geoscience and remote sensing symposium
45. Texas Water Development Board (2009) http://www.twdb.state.tx.us/home/index.asp
46. U.S. Environmental Protection Agency (2009) http://www.epa.gov/
47. Wang W, Yang J, Muntz RR (1997) STING: a statistical information grid approach to spatial data mining. In: Twenty-third international conference on very large data bases. Morgan Kaufmann, Athens, pp 186–195
48. World Health Organization (2009) http://www.who.int/

Wei Ding is an Assistant Professor of Computer Science at the University of Massachusetts Boston. She received her M.Sc. degree in Software Engineering from the George Mason University in 2000 and her Ph.D. degree in Computer Science from the University of Houston in 2008. She has 7-year teaching experience in Computer Science and 8-year industrial working experience in banking, software development, and web technology. Her main research interests include Data Mining, Machine Learning, Artificial Intelligence, Computational Semantics, and with applications to astronomy, geosciences, and environmental sciences. Her research projects are currently sponsored by NASA.

Christoph F. Eick received a Ph.D. in Computer Science from the University of Karlsruhe in Germany. He is an Associate Professor in the Department of Computer Science at the University of Houston. His areas of expertise include spatial data mining, clustering, machine learning, evolutionary computing, databases, and artificial intelligence. He published more than 100 papers in these areas. His current research centers on region discovery in spatial datasets, on the design of clustering algorithms that provide plug-in fitness functions, collocation mining, distance function learning, regional regression, and on the application of data mining to challenging problems in medicine, astronomy, geology, and environmental sciences. He also serves in program committees of major data mining conferences.

Xiaojing Yuan has years' experiences in embedding intelligence into sensors and actuators to deal with uncertainties. She authored and co-authored more than 50 technical papers; has one patent; filed two patents in USA in the last couple of years. She has been very active in professional organization: served as chair and co-chair in conferences and conference sessions for intelligent sensor networks, biomedical imaging processing and reviewed papers for high quality journals such as pattern recognition, wireless communication, IEEE Transaction on mobile computing, and information fusion. Since she joined the University of Houston, she established ISGRIN (Intelligent Sensor Grid and Informatics, www.tech.uh.edu/isgrin/) research lab, and developing enabling technology for wireless sensor network, new pattern recognition and modeling methods for biomedical and geospatial systems. Her research projects are funded by NSF, NASA, EIH, TWC, THI, and UH grants (UH new faculty grant, GEAR, and FDIP). Her paper was awarded best application paper in ICDM 2006 and the submission from her group (bmisgrin) scored 6th in the PAKDD 2009 data mining competition.

Jing Wang has a M.Sc in Computer Science From University of Houston. She is currently working at AOL Advertising as a software engineer. Her work at AOL include prototyping and developing predictive modeling, cutting edge optimization and data mining software for online advertising.

Jean-Philippe Nicot has a Ph.D. in Civil Engineering from the University of Texas at Austin. His research interests are diverse and include subsurface contaminant transport, in particular of natural contamination owing to high concentrations in arsenic, radionuclides, and other trace elements.