# On Regional Association Rule Scoping

Wei Ding[1], Christoph F. Eick[1], Xiaojing Yuan[2], Jing Wang[1], Jean-Philippe Nicot[3] *

## Abstract

*A special challenge for spatial data mining is that information is not distributed uniformly in spatial data sets. Consequently, the discovery of regional knowledge is of fundamental importance. Unfortunately, regional patterns frequently fail to be discovered due to insufficient global confidence and/or support in traditional association rule mining. Regional association rules, by definition, only hold in a subspace but not in the global space. One novel challenge is how to evaluate the impact of regional association rules. This paper centers on regional association rule scoping. We introduce a reward-based region discovery framework that employs clustering to find places where regional association rules are valid. We evaluate our approach in a real-world case study to discover arsenic risk zones in the Texas water supply. The experimental results are validated by domain experts and compared with published results on arsenic contamination.*

## 1 Introduction

The goal of spatial data mining is to automate the extraction of interesting, useful but implicit spatial patterns [10, 12, 16]. One special challenge in spatial data mining is that information is usually not uniformly distributed in spatial data sets. It has been pointed out in literature 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*"[8, 13]. Therefore, it is not surprising that domain experts are most interested in discovering hidden patterns at a regional scale rather than at a global scale.

*1. Computer Science Department, University of Houston, Houston, TX 77004, {wding,ceick,jwang29}@uh.edu
2. Engineering Technology Department, University of Houston, Houston, TX 77004,xyuan@uh.edu
3. Bureau of Economic Geology, John A. & Katherine G. Jackson School of Geosciences, University of Texas at Austin, jp.nicot@beg.utexas.edu

Unfortunately, when using traditional association rule mining, regional patterns frequently fail to be discovered due to insufficient global confidence and/or support. A common approach to alleviate the problem is using a small support threshold, but this approach usually suffers from a combinatorial explosion in the number of rules generated. Selecting a proper confidence threshold is even more subtle. Let's consider an association rule that suggests a well up to 251.5-feet deep is associated with dangerous arsenic levels:

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

Assuming that the minimum confidence threshold is 70%, the pattern is not strong enough to be identified globally in Zone A and Zone B (see Table 1) because its confidence, $\frac{1000}{2000} = 50\%$, is less than the minimum confidence threshold. But in Zone A the rule holds because its confidence, $\frac{400}{500} = 80\%$, is above the threshold. Notice that this rule does not hold in Zone B, due to its low confidence ($\frac{600}{1500} = 40\%$). This reversal of the direction of an association in the global dataset is also known as Simpson's Paradox in statistics [6].

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. This fact leads to novel challenges for regional association mining: how to determine regions from which regional association rules will be derived and how to evaluate the impact of regional association rules. The first problem has been addressed in our previous work in [5]. The goal of this paper is to address the second issue by utilizing the duality between regional patterns and regions where the patterns are supported: regions are used to discover regional association rules, and then regional association rules are used to determine regions in which the association rules are valid. Such regions provide a quantitative measure of how significant a regional association rule is in the global space.

This paper proposes an approach that computes the scope of regional association rules by employing a clustering algorithm in such a way as to maximize an externally given fitness function. The fitness function is

| Well Depth | Dangerous | Safe | Total |
|---|---|---|---|
| (0, 251.5] | **1000** | 1000 | **2000** |
| (251.5, $\infty$) | 1200 | 800 | 2000 |
| Total | 2200 | 1800 | 4000 |

| | Well Depth | Dangerous | Safe | Total |
|---|---|---|---|---|
| ZoneA | (0, 251.5] | **400** | 100 | **500** |
| | (251.5, $\infty$) | 1050 | 450 | 1500 |
| ZoneB | (0, 251.5] | **600** | 900 | **1500** |
| | (251.5, $\infty$) | 150 | 350 | 500 |
| Total | | 2200 | 1800 | 4000 |

**Table 1. Contingency tables between well depth and arsenic concentration.**

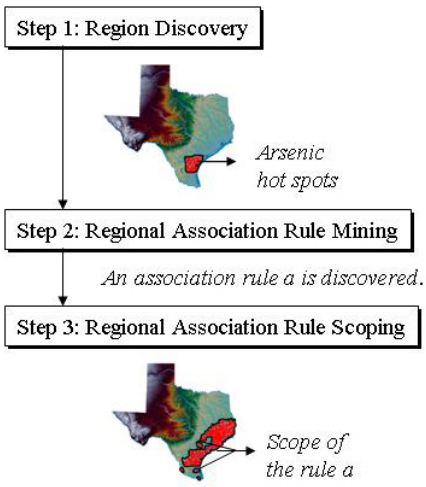


**Figure 1. An example Regional Association Rule Scoping.**

especially constructed to encode the preferences of a domain expert. Each cluster is assigned a "reward" value. A cluster receives higher reward if a regional association rule exhibits stronger confidence and support. We empirically evaluate the effectiveness of our method using a real-world dataset describing the arsenic ground water pollution in Texas, a problem for which the identification of causality of arsenic contamination is of great interest to domain scientists. Our experimental results not only confirm and validate research results in geosciences, and also lead to the discovery of novel findings that need to be further studied by domain experts. Figure 1 illustrates the basic procedure of our approach. An association rule $a$, *the wells with nitrate concentration lower than 0.085mg/l have dangerous arsenic concentration level*, is discovered from an arsenic hot spot area in South Texas with 100% confidence. The scope of the association rule $a$ is a much larger area which mostly overlaps with the Texas Gulf Coast. Statistical analysis shows that the rule $a$ cannot be discovered at Texas state level due to its insufficient confidence (less than 50%).

**Related Work.** To our best knowledge, no previous work has been done in spatial data mining to determine the scope of association rules. The areas most relevant to our work are spatial association rule mining [10], co-location mining [17], and localized association rule mining [2].

Spatial association rule mining [10] extends association rule mining to spatial data sets. It denotes association relationships among a set of predicates, where there exists at least one spatial predicate. Co-location mining [17] identifies a subset of boolean spatial features whose instances are frequently located together in close proximity. Both approaches focus on finding frequent, global patterns that characterize the complete dataset, whereas our approach centers on finding interesting places where associations of a subset of non-spatial features hold. Localized association rule mining [2] is a similar approach as ours: it discovers association rules that hold in local segments of the basket data that are determined using clustering, but their discovery is limited to non-spatial basket datasets.

**Contributions.** We define and address a previously unstudied problem of regional association rule scoping. We propose a unique reward-based region discovery framework that employs clustering to find interesting subspaces for regional association rules. We evaluate our method in a real-world case study that identifies interesting associations and places in Texas with respect to arsenic risk patterns.

## 2 An Integrated Approach for Regional Association Rule Scoping

### 2.1 Measuring the Interestingness of a Set of Regions

Our region discovery method employs a reward-based evaluation scheme that evaluates the quality of the generated regions. Let $\mathbb{D}$ be a spatial dataset, and $S = \{s_1, s_2, \ldots, s_l\}$ be a set of spatial attributes, such as longitude or latitude; $A = \{a_1, a_2, \ldots, a_m\}$ be a set of non-spatial attributes (continuous attributes are required to be transformed into nominal attributes); let

$$\begin{aligned} I &= S \cup A \\ &= \{s_1, s_2, \ldots, s_l, a_1, a_2, \ldots, a_m\} \end{aligned}$$

be the set of all the items in $\mathbb{D}$. Given a set of regions $R = \{r_1, \ldots, r_n\}$, the fitness of $R$ is defined as the

sum of the rewards obtained from each region $r_i$ ($i = 1 \ldots n$).

$$q(R) = \sum_{i=1}^{n} (i(r_i) \times |r_i|^{\beta}) \quad (1)$$

where $i(r_i)$ is the interestingness measure of region $r_i$ with respect to a given association rule. $|r_i|^{\beta}$ ($\beta > 1$) in $q(R)$ increases the value of the fitness nonlinearly with respect to the region size $|r_i|$. The amount of premium put on the size of the region is controlled by the user-determined value of parameter $\beta$. The evaluation scheme encourages the merging of regions if their overall interestingness does not decrease.

We use a clustering algorithm to seek for a set of clusters (regions) such that the sum of rewards over all of its constituent regions is maximized. A region is identified as a cluster that receives a high reward. It is a contiguous subspace that contains a set of spatial objects. 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 $r_1, \ldots, r_n$ such that:

1. $r_i \cap r_j = \emptyset, i \neq j$. The regions are disjoint.
2. $R = \{r_1, \ldots, r_n\}$ maximizes $q(R)$.
3. $r_1 \cup \ldots \cup r_n \subseteq \mathbb{D}$. The generated regions are not required to be exhaustive with respect to the global dataset $\mathbb{D}$.
4. $r_1, \ldots, r_n$ are ranked based on the reward values. Regions that receive no reward are discarded as outliers.

## 2.2 Association Rule Scoping

The goal of association rule scoping is to compute a set of regions where a given association rule is valid. Let $a$ be an association rule, $r$ be a region, $conf(a, r)$ denotes the confidence of $a$ in region r, and $sup(a, r)$ denotes the support of $a$ in $r$.

**Definition 1.** The *scope* of an association rule $a$ is the regions where the association rule $a$ satisfies the $min\_sup$ and $min\_conf$ thresholds, where $min\_sup$ and $min\_conf$ are the corresponding support and confidence thresholds.

In fact, the scope of a regional association rule represents the spatial impact of this regional pattern.

We define the interestingness, $i(r)$, of region $r$ with respect to a given association rule $a$ as follows:

$$i(r) = \begin{cases} 0, & \text{if } sup(a,r) < min\_sup \times \delta_1 \text{ or } \\ & conf(a,r) < min\_conf \times \delta_2, \\ (\frac{sup(a,r)}{min\_sup})^{\eta_1} (\frac{conf(a,r) - min\_conf \times \delta_2}{1 - min\_conf \times \delta_2})^{\eta_2}, & \text{otherwise.} \end{cases} \quad (2)$$

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 Equation 2, the threshold $min\_sup \times \delta_1$ 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 of low interestingness. Values of parameters $\eta_1$ and $\eta_2$ ($\eta_1, \eta_2 > 0$) determine the weight to the increment of the support and confidence respectively.

The measure of interestingness is designed to efficiently identify the scope of a given regional association rule. Firstly, in contrast to traditional association rule mining, the proposed measure of interestingness uses "soft" instead of "hard" thresholds to avoid a crisp effect [4]. For example, with $\delta_1 = \delta_2 = 0.9$, it rewards regions as long as their confidence or support thresholds are within 90% of the hard thresholds $min\_conf$ and $min\_sup$. Assume $min\_sup = 10\%$, $min\_conf = 80\%$, and an association rule whose support is 9% and confidence is 100% in a region $r'$. Instead of assigning zero reward to the region $r'$, we argue to reward the region because the confidence of the rule is significantly above $min\_conf$ threshold and its support is just a little bit lower (1%) than the $min\_sup$ threshold. Secondly, our approach uses a quantitative evaluation method that assigns higher degree of interestingness and consequently higher reward to regions whose support and confidence are high with respect to an association rule of interest. Thirdly, once an association rule $a$ is discovered from a particular region $r$, we already know that the region $r$ from which $a$ is originated, receives a positive reward due to the fact that $a$ satisfies the support and confidence threshold in $r$.

## 2.3 Other Applications of Regional Association Rule Scoping

Association rule scoping has many applications that lie outside of the regional association mining methodology we just introduced. First it is important to emphasis that our approach 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. Second, a domain expert may be interested to explore how the scope of an association rule changes, if an association rule is 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 from the association rule scoping can serve as a source again for creating new interesting 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.

### 2.4 Clustering Algorithm

In our regional association rule scoping framework, different interestingness functions correspondent to various domain interests can easily be plugged into a clustering algorithm. We have used the clustering algorithm SCMRG (Supervised Clustering using Multi-Resolution Grids) on hot spot discovery in our previous work [7]. In this paper, we revise the algorithm using the fitness defined in Equation 1 and 2 for regional association rule scoping. SCMRG is a hierarchical, grid-based method that utilizes a divisive, 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 if the sum of the rewards of the lower level cells is not decreased. A cell is partitioned further only if it improves its fitness at a lower level of resolution. A queue data structure is used to store all the cells that need to be processed. The algorithm traverses through the hierarchical structure and examines those cells in the queue from the higher level. Finally, the algorithm collects all the cells that have been labeled as clusters from different levels, and neighbor clusters are merged if fitness can be improved.

## 3 A Real-World Case Study: Arsenic Risk Zones in Texas

We evaluate our regional association rule scoping method using an arsenic water pollution dataset. Approximately 6% of the Texas wells are in violation with the new EPA (Environment Protection Agency) arsenic maximum contaminant limit (MCL) for drinking water [15]. This raises concerns about how to prepare the Texas communities to cope with the threat of arsenic contamination in their water supply. The arsenic dataset consists of 24 attributes of 12055 water wells collected from the Texas Ground Water Database [18]. Of the 24 attributes, 4 spatial attributes ($S$) are latitude and longitude, river basin, and state zone; 19 non-spatial attributes ($A$) are well depth, concentration of fluoride, nitrate, and other chemical metal elements selected by domain experts [9, 11, 14]; finally 1 arsenic class attribute indicates *safe* or *dangerous* wells. A well is dangerous if its arsenic concentration level is above the MCL of $10\mu g/l$.

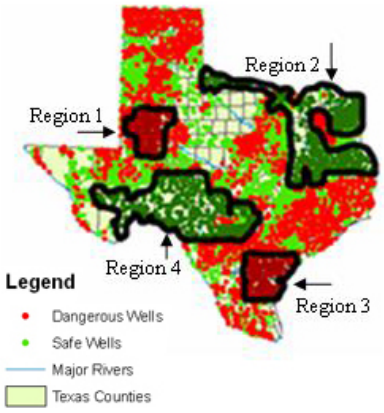


**Figure 2. Interesting regions are identified by SCMRG**

### 3.1 Experimental Evaluation

We have implemented the clustering algorithm SCMRG using two different functions of interestingness for hot spots/cool spots discovery (for details see [7]) and for regional association rule scoping (using Equation 2). Hot spots/cool spots are regions in which the density of the class of interest is much higher/lower than its average density in the whole dataset. The SCMRG algorithm uses spatial attributes longitude, latitude, and arsenic class attribute to search for hot spots/cool spots of arsenic. Figure 2 shows the result of such a run, where safe wells are in green (or light grey), dangerous wells in red (or dark grey). It illustrates four most highly rewarded regions – Region 1 and 3 are regions of hot spots (high density of dangerous wells), and Region 2 and 4 are regions of cool spots (high density of safe wells). Region 1, southern half of the High Plains, and Region 3, the south Gulf

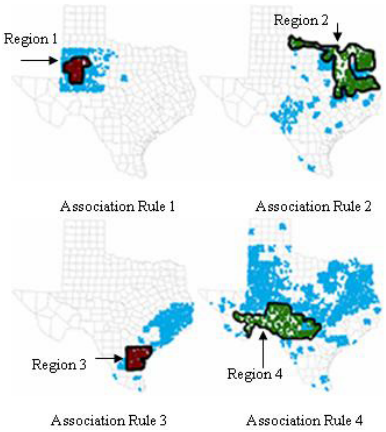


**Figure 3. Region - Regional association rule - Scope. Legend: regions are highlighted by bold border line; scopes are in color blue (or light grey).** $\beta = 1.01, \eta_1 = 1, \eta_2 = 1.1, \delta_1 = \delta_2 = 0.9, min\_sup = 10\%, min\_conf = 80\%$

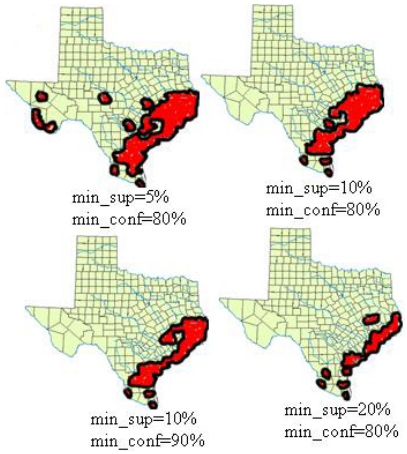


**Figure 4. The scope of a particular rule changes based on the different values of** $min\_sup$ **and** $min\_conf$**.** $\beta = 1.01, \eta_1 = 1, \eta_2 = 1.1, \delta_1 = \delta_2 = 0.9, min\_sup = 10\%, min\_conf = 80\%$

Coast, overlap with the arsenic risk zone discussed by geoscientists in [15, 1, 14].

In the next step, regional association rules are generated from each region. We extend the Apriori algorithm [3] to generate association rules that are related with arsenic class attribute: the algorithm not only prunes infrequent candidate itemsets, but also discards itemsets that do not contain the arsenic class labels. The algorithm uses the whole set of attributes excluding longitude and latitude, with $min\_sup = 10\%$ and $min\_conf = 80\%$. The following four regional association rules with confidence 100% are discovered from Region 1, 2, 3, and 4 respectively. Association rules 1 and 3 are confirmed in arsenic literature [9, 11].

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

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

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

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

Finally we seek for scope of those interesting regional association rules. Figure 3 depicts the scope of the above 4 association rules. 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 ($< 251.5$ feet) indicate the aquifer is thin, thus nitrate comes from surface contamination ($> 28.31\ MG/L$), and 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 is similar. The scope of association rule 2 and 4 represent the areas where arsenic contamination is low. They are interesting places that domain scientists will future explore. The experimental results also confirm our discussion in Section 2.2: the region, where an association rule is originated, is a subset of the scope where the association rule holds.

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 Figure 3 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 association rules are discovered globally. We can also fine tune the measurement interestingness in association rule scoping by varying its support and confidence thresholds for a given association rule. Figure 4 shows such a

scope tuning for the association rule 3. Typically, lower value of $min\_sup$ results in larger scope, higher value of $min\_conf$ results in smaller scope.

Our SCMRG algorithm is computationally efficient. On average, it takes 3.031 seconds for hot spots/cool spots discovery, and 4.68 seconds for regional association rule scoping. The computer uses Intel(R) Pentium(R) M, CPU 1.2GHz, 632 MB of RAM.

## 4 Discussion and Future Work

One critical requirement for spatial data mining is to analyze data at different levels of granularity. We define and address a novel problem of regional association rule scoping. Scope is the fundamental property of regional association rules. Moreover, rule scoping provides unique capabilities to domain experts to explore the impact of changing an association rule versus the regions it applies and then identify the root cause of it. We have evaluated the proposed framework in a real-world case study that centers on identifying spatial risk zones of arsenic in Texas water supply. Our approach was able to re-discover hypotheses that already have been established in the scientific literature on arsenic contamination, and also discovered several novel hypotheses concerning the causes of arsenic contamination that deserve further exploration.

An interactive tool of rule scoping that allows for changing existing rules and for testing new rules is under development so that the changes in rule scope can be visually presented to a domain expert, as illustrated in Figure 4. Such capabilities undoubtedly are very useful for exploratory hypothesis testing for a domain expert. It will not only increase the understanding of spatial data sets itself, but also identify the scope for further studies such as characterization, pattern recognition, and modeling of the particular regions.

## References


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

[2] C. C. Aggarwal, C. M. Procopiuc, and P. S. Yu. Finding localized associations in market basket data. *IEEE Transactions on Knowledge and Data Engineering*, 14:51–62, 2002.

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

[4] S. Bistarelli and F. Bonchi. 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*, volume 3721, Porto, Portugal, October 2005. Springer.

[5] W. Ding, C. F. Eick, J. Wang, and X. Yuan. A framework for regional association rule mining in spatial datasets. In *The 6th IEEE International Conference on Data Mining (ICDM)*, Dec. 2006.

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

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

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

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

[10] K. Koperski and J. 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.

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

[12] R. Munro, S. Chawla, and P. Sun. Complex spatial relationships. In *The Third IEEE International Conference on Data Mining (ICDM)*, 2003.

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

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

[15] B. R. Scanlon and J. P. Nicot. Evaluation of arsenic contamination in texas. Technical report, final report prepared for Texas Commission on Environmental Quality, under contract no. UT-08-5-70828, 2005.

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

[17] S. Shekhar and Y. Huang. Discovering spatial colocation patterns: A summary of results. *Lecture Notes in Computer Science*, 2121, 2001.

[18] Texas Water Development Board. *http://www.twdb.state.tx.us/home/index.asp*, 2007.