Finding Regional Co-location Patterns for Sets of Continuous Variables in Spatial Datasets

Christoph F. Eick, Rachana Parmar, Wei Ding†
Department of Computer Science
University of Houston
Houston, TX 77204-3010

Tomasz F. Stepinski‡
Lunar and Planetary Institute
Houston, TX 77058

Jean-Philippe Nicot*
Bureau of Economic Geology,
Jackson School of Geosciences
University of Texas at Austin
Austin, TX 78712

ABSTRACT

This paper proposes a novel framework for mining regional co-location patterns with respect to sets of continuous variables in spatial datasets. The goal is to identify regions in which multiple continuous variables with values from the wings of their statistical distribution are co-located. A co-location mining framework is introduced that operates in the continuous domain without the need for discretization and which views regional co-location mining as a clustering problem in which an externally given fitness function has to be maximized. Interestingness of co-location patterns is assessed using products of z-scores of the relevant continuous variables. The proposed framework is evaluated by a domain expert in a case study that analyzes Arsenic contamination in Texas water wells centering on regional co-location patterns. Our approach is able to identify known and unknown regional co-location patterns, and different sets of algorithm parameters lead to the characterization of Arsenic distribution at different scales. Moreover, inconsistent co-location sets are found for regions in South Texas and West Texas that can be clearly attributed to geological differences in the two regions, emphasizing the need for regional co-location mining techniques. Moreover, a novel, prototype-based region discovery algorithm named CLEVER is introduced that uses randomized hill climbing, and searches a variable number of clusters and larger neighborhood sizes.

Keywords

spatial data mining, regional co-location mining, regional knowledge discovery, clustering, finding associations between continuous variables.

1. INTRODUCTION

As the ability to capture and store information expands, spatial context has emerged as an increasingly important part of discovering knowledge in large amounts of data. The motivation for regional knowledge discovery is driven by the fact that global statistics seldom provide useful insight and that most relationships in spatial datasets are geographically regional. The need for robust tools capable of extracting knowledge from large spatial datasets is critical for advancing scientific research in areas ranging from global climate change and its effect on regional ecosystems, to environmental risk assessment and for choosing appropriate environmental policies.

Discovery of co-location patterns, a co-occurrence of different types of features at approximately the same locations, is an important example of a data mining task with many practical applications. Most existing research has concentrated on discovering global co-location patterns with respect to categorical features, which identify sets of classes whose instances co-occur in geographical proximity with high frequency. A classic example [23] of such a relationship is the co-location of two types of animals, the Nile crocodile and the Egyptian plover, which is traced by domain scientists to their symbiotic relationship.

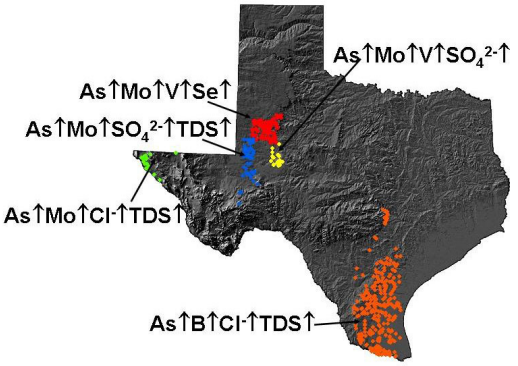

Figure 1. Regional co-location patterns involving chemical concentrations in Texas wells.

However, not all real-life problems are susceptive to the categorical formulation. In a broad range of problems the spatial dataset is given in the form of continuous variables. Formulating such a problem in terms of categorical, discrete features is not natural. In this paper, we are interested in identifying regions where extreme values of different continuous variables are present in geographical proximity. Figure 1 illustrates what we are trying to accomplish for a dataset that contains concentrations of chemicals in different wells in Texas. The goal is to find regions (sets of wells that cover a contiguous area) in which concentrations of multiple chemicals take extreme values. The figure shows the result of running a regional co-location mining algorithm on such a dataset; it identifies five regions with five different patterns of chemical concentrations. For example, two interesting regions related with high Arsenic concentrations are identified in the Western half of Texas: the first region, colored in red, contains high concentrations of Arsenic (As↑), Molybdenum (Mo↑), Vanadium (V↑) and Selenium (Se↑), whereas the second region, colored in blue, is characterized by high concentrations of Arsenic (As↑), Molybdenum (Mo↑), Sulfate (SO_4^{2-}↑) and Total Dissolved Solids (TDS↑).

† {ceick, rparmar, wding}@uh.edu

‡ tstepinski@lpi.usra.edu

* jp.nicot@beg.utexas.edu

In this paper, we propose and evaluate a novel framework for discovering co-location regions and their associated patterns in a highly automated fashion in continuous datasets without the need for discretization. The proposed framework treats region discovery as a clustering problem in which clusters have to be obtained that maximize an externally given fitness function. The fitness function combines contributions of interestingness from each individual cluster and can be customized by a domain expert. The framework allows the actual clustering task to be performed by a variety of different algorithms. A highly desirable feature of our approach is its search-engine-like capabilities, returning a set of regions ranked by interestingness thus providing a domain expert with pertinent information.

Related Work. The relevant research spans three areas:

Hot Spot Discovery in Spatial Statistics. In [4] the detection of hot spots using a variable resolution approach was investigated in order to minimize the effects of spatial superposition. The definition of hot spots was extended in [18] using circular zones for multiple variables. [13, 21] propose a popular method to find hot spots in spatial datasets relying on the G* Statistic. The G* Statistic detects local pockets of spatial association. 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 and cold spots. However, it should be noted that such aggregates are not formally defined clusters, as the G*-approach has no built-in clustering capabilities.

Spatial Co-Location Pattern Discovery. Shekhar *et al.* discuss several interesting approaches to mine co-location patterns, which are subsets of Boolean spatial features whose instances are frequently located together in close proximity [23, 28, 29]. Huang *et al.* proposed co-location mining involving rare events [14]. In [15], Huang and Zhang explored the relations between clustering and co-location mining. Instead of clustering spatial objects, the features of spatial objects are clustered using a proximity function that is designed to find co-locations. However, it should be stressed that all the approaches mention above are restricted to categorical datasets and center on finding global co-location patterns, whose scope is the whole dataset. Our approach, on the other hand, as we will explain later in more detail, centers on discovering regions and regional co-location patterns whose scope is a subspace of the whole dataset. Localized association rule mining [2] takes a similar approach to ours, but it discovers association rules that hold in local clustered basket data. Thus, their discovery is limited to non-spatial basket datasets.

Finding Associations between Continuous Attributes. Most of the approaches to mine association rules in datasets containing continuous attributes use discretization. In [26], numerical attributes are discretized and then adjacent partitions are combined as necessary. This leads to information loss and can generate spurious rules. Aumann et al. [3] introduce numerical association rules that support statistical predicates for continuous attributes, such as variance, and algorithms that mine such rules. In [5], rank correlation is used to mine associations between numerical attributes. Basically, continuous attributes are transformed to ordinal attributes, and a method is proposed to find sets of numerical attributes with high attribute value associations. Achtert [1] and Jaroszewicz [16] propose different methods for deriving equations describing relationships between continuous variables in datasets.

Contributions.

1. A novel measure of interestingness and a regional co-location mining framework is proposed that identifies places in which continuous variables taking values from the wings of their respective distributions co-occur. The proposed method directly operates in the continuous space without any need for discretization.
2. Techniques are proposed that find regional and not global associations between continuous variables. One particular challenge of this task is that the employed algorithms need to search for both interesting places and interesting patterns at the same time.
3. We apply our framework to the problem of identifying regional co-location patterns with respect to high and low Arsenic concentrations in Texas water supply. A thorough analysis of this case study is presented including comparison of results obtained using different parameter settings and an assessment of the found patterns by a domain expert is given.
4. As by product, a novel prototype-based clustering algorithm named *CLEVER* is introduced which employs randomized hill climbing, allows for a variable number of clusters, and searches larger neighborhood sizes to battle premature convergence.

2. REGION DISCOVERY FRAMEWORK

As mentioned in the previous section, we are interested in the development of frameworks and algorithms that find interesting regions in spatial and spatio-temporal datasets. The presented framework has originally been introduced in [10, 11], and will be generalized in this section to mine spatio-temporal datasets that contain multiple continuous variables. A novel measure of interestingness for mining co-locations involving continuous attributes that is embedded into this framework will be introduced in Section 3.

Our work assumes that region discovery algorithms we develop operate on datasets containing objects $o_1,..,o_n$: $O=\{o_1,..,o_n\}\in F$ where F is relational database schema and the objects belonging to O are tuples that are characterized by attributes $S\cup N$, where:

$S=\{S_1,\ldots,S_p\}$ is a set of spatial and temporal attributes.

$N=\{A_1,\ldots,A_q\}$ is a set of other, non-geo-referenced attributes.

Dom(S) and Dom(N) describe the possible values the attributes in S and N can take; that is, each object $o\in O$ is characterized by a single tuple that takes values in $F=Dom(S)\times Dom(N)$. Datasets that have the structure we just introduced are called *geo-referenced datasets* in the following, and O is assumed to be a geo-referenced dataset throughout this paper. The purpose of the framework is to find interesting places, called regions in the following, in geo-referenced datasets. Regions are assumed to be contiguous areas in the spatial-temporal space Dom(S) which is a subspace of F. A region has an extension which is the set of objects in O it contains and an intension that describes the area it occupies. For example, the intension of the orange region in Fig. 1 is South Texas, and its extension includes the water wells in O that are located in this region.

The region discovery framework employs additive, plug-in fitness functions q that capture what kind of regions are of interest to the

domain expert; moreover, fitness functions are assumed to have the following structure:

$$q(X) = \sum_{c \in X} reward(c) = \sum_{c \in X} i(c) * |c|^{\beta} \quad (1)$$

where i(c) denotes the interestingness of region c—a quantity to reflect a degree to which regions are "newsworthy". It is important to find regions at different levels of granularity. The amount of premium put on the size of the extension of a region ('|c|' denotes the cardinality of c) is controlled by the value of parameter β. A region reward is proportional to its interestingness, but rewards increase with region size non-linearly (β>1).

Given a geo-referenced dataset O, there are many possible algorithms to seek interesting regions in O with respect to a plug-in fitness function q, subject to the following specification:

Given: O, q, and possibly other input parameters
Find: X={r_1,...,r_k} that maximize q({r_1,...,r_k}) subject to the following constraints:
(1) $r_i \subseteq O$ (i=1,…,k)
(2) r_1 , r_2, …, r_k are contiguous[1] in Dom(S)
(3) $r_i \cap r_j = \emptyset$ (i≠j)

So far, nine region discovery algorithms (four representative-based, three agglomerative, one divisive, and one density-based region discovery algorithm) have already been designed and implemented in our past work [6, 9, 12]. A novel unpublished, prototype-based clustering algorithm named CLEVER will be later used to evaluate the presented co-location mining approach; therefore, CLEVER will be briefly described in this section.

Prototype-based clustering algorithms construct clusters by seeking a set of "optimal" representatives; clusters are then created by assigning objects in the dataset to the closest representative. Popular prototype-based clustering algorithms are K-Medoids/PAM [17] and K-means [19]. CLEVER (CLustEring using representatiVEs and Randomized hill climbing) seeks to maximize the fitness function q(X). The algorithm (see Figure 2) starts with randomly selecting k' representatives from O—k' is a parameter of the algorithm. It samples p solutions in the neighborhood of the current solution; unlike CLARANS [20] which picks the first best neighbor as the next solution, CLEVER evaluates all the p neighbors and picks the best among them. Neighboring solutions of the current solution are created using three operators: 'Insert' – inserts a new representative into the current solution, 'Delete' – deletes a representative from the current solution and 'Replace' – replaces a representative with a non-representative. Each operator has a certain selection probability and representatives to be manipulated are chosen at random. The algorithm also allows for larger neighborhood sizes; the experiments in this paper were run for neighborhood size 3: in this case, solutions that are sampled are generated by applying three randomly selected operators to the current solution. Moreover, to battle premature convergence, CLEVER re-samples p'>p solutions before terminating. Figure 2 gives the pseudo-code for CLEVER.

CLEVER

Inputs: k', neighborhood-size, p, p'

Outputs: regions, region representatives, number of representatives (k), fitness, interestingness,…

Algorithm:

1. Create a current solution by randomly selecting k' representatives from O.
2. Create p neighbors of the current solution randomly using the given neighborhood definition.
3. If the best neighbor improves the fitness, it becomes the current solution. Go back to step 2.
4. If the fitness does not improve, the solution neighborhood is re-sampled by generating q more neighbors. If re-sampling does not improve the current solution, terminate; otherwise, go back to step 2 replacing the current solution by the best solution found by re-sampling.

Figure 2. Pseudo-code of algorithm CLEVER

By adding and deleting representatives and by using neighborhood size of larger than one, CLEVER samples from much larger neighborhood of the current solution. This characteristic distinguishes CLEVER from other prototype-based clustering algorithms.

3. A MEASURE OF INTERESTINGNESS FOR REGIONAL CO-LOCATION PATTERNS

In the following a function i is introduced that measures the interestingness of co-location patterns for a region c. The pattern $A\uparrow$ denotes that attribute A has high values and the pattern $A\downarrow$ indicates that attribute A has low values. For example, the pattern {$A\uparrow$, $B\downarrow$, $D\uparrow$} describes that high values of A are co-located with low values of B and high values of D.

Let

O be a dataset

$c \subseteq O$ be a region

$o \in O$ be an object in the dataset O

$N=\{A_1,\ldots,A_q\}$ be the set of non-geo-referenced continuous attributes in the dataset O

$Q=\{A_1\uparrow, A_1\downarrow,\ldots, A_q\uparrow, A_q\downarrow\}$ be the set of possible base co-location patterns

$B \subseteq Q$ be a set of co-location patterns

P(B) be a predicate over B that restricts the co-location sets considered[2]

z-score(A,o) be the z-score[3] of object o's value of attribute A

[1] For each pair of objects belonging to the same region there has to be a path connecting the two objects that solely traverses this region.

[2] e.g. P(B)=|B|<5 ("only co-locations sets with cardinalities 2, 3 and 4 are considered") or P(B)=$As\uparrow \in B$ ("only look for patterns involving high arsenic")

$$z(A\uparrow, o) = \begin{cases} z\text{-}score(A,o) \text{ if } z\text{-}score(A,o) > 0 \\ 0 \quad otherwise \end{cases} \quad (2)$$

$$z(A\downarrow, o) = \begin{cases} -z\text{-}score(A,o) \text{ if } z\text{-}score(A,o) < 0 \\ 0 \quad otherwise \end{cases} \quad (3)$$

z(p,o) is called the *z-value* of base pattern p∈Q for object o in the following. The interestingness of an object o with respect to a co-location set B⊆Q is measured as the product of the z-values of the patterns in the set B. It is defined as follows:

$$i(B,o) = \prod_{p \in B} z(p,o) \quad (4)$$

When using the above formula, the more extreme the z-values of the involved objects are, the bigger the above product becomes—moreover, if the value of a continuous variable does not match its suggested pattern in B its z-value is 0 and the interestingness is therefore 0 as well. Although this approach compresses multiple z-values into a single value, the product of z-values still allows for meaningful statistical interpretation using the geometric mean; for example, if the geometric mean of the z-values of the patterns in set B is 1.5, this suggests that values of the involved variables are at an average 1.5 standard deviations off their mean value.

In general, the interestingness of a region can be straightforwardly computed by taking the average interestingness of the objects belonging to a region. However, using this approach some very large products might dominate interestingness computations. For some domain experts just finding a few objects with very high products in close proximity of each other is important, even if the remaining objects in the region deviate from the observed pattern. In other cases, domain experts are more interested in patterns with highly regular products so that all or almost all objects in a region share this pattern, and are less interested in a few very high products. To satisfy the needs of both groups, our approach additionally considers purity when computing region interestingness, where *purity(B,c)* denotes the *percentage of objects o∈c for which i(B,o)>0.* In summary, the interestingness of a region c with respect to a co-location set B, denoted by *φ(B,c)*, is computed as follows:

$$\varphi(B,c) = \frac{\left(\sum_{o \in c} i(B,o)\right)}{|c|} * purity\,(B,c)^{\theta} \quad (5)$$

The parameter θ∈[0,∞) controls the importance attached to purity in interestingness computations; θ=0 implies that purity is ignored, and using larger values increases the importance of purity.

The un-normalized, *raw interestingness* of a region c, denoted by $\kappa_S(c)$ is measured as the maximum interestingness *φ(B,c)* observed over all subsets B⊆Q with cardinalities 2 and higher considered, subject to the restrictions imposed in predicate P.

$$\kappa_S(c) = \max_{B \subseteq Q \& |B|>1 \& P(B)} \varphi(B,c) \quad (6)$$

The normalized[4] interestingness of a region c, *i*(c), is defined as follows:

$$i(c) = \begin{cases} (\kappa_S(c) - th)^{\eta} \text{ if } \kappa_S(c) > th \\ 0 \quad otherwise \end{cases} \quad (7)$$

The threshold parameter *th≥0* is introduced to weed out regions c with $\kappa_S(c)$ close to 0. Moreover, η is a scaling factor that allows modifying raw interestingness super-linearly by choosing η>1, and sub-linearly by choosing η<1.

Finally, as discussed earlier, the reward of the region c is computed as follows:

$$reward(c) = i(c) * |c|^{\beta} \quad (8)$$

Within our present focus, *i*(c) must encapsulate a degree to which extreme values of variables are present together in region c. The region size is denoted by |c|, and the quantity $i(c)*|c|^{\beta}$ can be considered as a "reward'' given to a region c; we seek X such that the sum of rewards over all of its constituent regions is maximized. The amount of premium put on the size of the region is controlled by the value of parameter β. The purpose of β is to control the granularity of the regions that are discovered; regions with larger extensions tend to be discovered using higher values of β.

Example. Table 1 shows the extension of an example region c, containing four objects with the indicated values for attributes C and D, and intermediate values used in calculating *i*(B,o) for pattern B={C↑, D↓}. Column 3 and 4 display the z-values for C↑ and D↓ respectively that are calculated using formulas (2) and (3). Column 6 displays i(B,o) as per formula (4). We can see that purity of pattern B is 0.5. Assuming θ=1, using formula (5) we obtain: φ(B,c)=((0.24+0.24)/4)*0.5=0.06.

Table 1. Interestingness computations for a region.

ID	C z-score	D z-score	C↑	D↓	*i*(B, o)
1	0.43	-0.56	0.43	0.56	0.24
2	0.42	-0.56	0.42	0.56	0.24
3	-0.06	0.13	0	0	0
4	-0.57	-0.22	0	0.22	0

It is not feasible to employ pruning using maximum valued co-location sets of size m when computing co-location sets of size (m+1), because the interestingness function *i* is not anti-monotone. Hence when computing *i*(c) all legal subsets B⊆Q with cardinality 2 and higher have to be considered. We explain why pruning is not feasible by giving a counter example. Let us assume that for a

[3] The z-score of value a for attribute A is: $(a-\mu_A)/\sigma_A$ where μ_A is the mean value and σ_A is standard deviation of attribute A.

[4] One assumption underlying our framework is that clusters that are not interesting for a domain expert receive a reward of 0. The framework treats objects belong to clusters that receive no reward as outliers. Therefore, fitness functions are usually normalized and scaled in collaboration with domain experts based on what the domain expert finds "newsworthy".

region c, B={$A_1\uparrow$,$A_2\uparrow$} is the binary pattern with the highest interestingness; however, the highest interestingness pattern of size three B' may not necessarily contain B as the subset. Let's assume, that all objects with positive z-scores for A_1 and A_2 in region c have zero z-scores for remaining attributes $A_3,\ldots,A_5$ and that there is at least one object in region c that has zero z-scores for A_1 and A_2 and positive z-scores for remaining attributes. All the patterns of size 3 having B as a subset will therefore have interestingness 0, but {$A_3\uparrow$,$A_4\uparrow$,$A_5\uparrow$}'s interestingness is above 0. Therefore, the maximum interestingness pattern of size 3 does not contain {$A_1\uparrow$,$A_2\uparrow$} as a subset for region c.

In general, when the co-location framework is used without imposing any restrictions on co-location sets considered a large number of co-location sets ($O(2^{|N|})$) has to be evaluated. This leads to a very slow performance of the co-location mining algorithm. Moreover, when a large number of disjoint co-location sets are searched in parallel, the crudeness of the maximum operation in formula 6 results that only the most interesting pattern will be reported, and other interesting patterns will be ignored.

Therefore, to alleviate this problem, when the framework is used in practice for medium sized or large attribute sets, it is mandatory to restrict pattern exploration by imposing constraints on co-location sets. A promising approach is to use seeded patterns; the idea here is to request that co-location sets have to contain certain patterns. For example, in the experiments that will be discussed in the next section, only co-location sets that contain either As↑ or As↓ are considered, restricting the number of patterns significantly. In the seeded approach, instead of finding all the patterns in a single run, we run the co-location mining algorithm multiple times with different seeds. On the positive side, this allows for a more focused and quicker discovery of co-location patterns; on the negative side, once seeded exploration is used, results of multiple runs have to be analyzed and integrated, representing a new challenge for co-location mining that will be revisited in Section 4.

4. CASE STUDY: FINDING REGIONAL CO-LOCATION PATTERNS WITH RESPECT TO ARSENIC IN THE TEXAS WATER SUPPLY

We evaluated our framework in a real world case study to discover regional co-location patterns involving Arsenic and other chemicals in the Texas water supply.

Dataset Description and Preprocessing. Datasets used in this case study are created using the Groundwater database (GWDB) maintained by the Texas Water Development Board [27]. Long term exposure to low level concentrations of Arsenic causes cancer [23]. Figure 3 shows various aquifers and Arsenic pollution sites on the map of Texas reported by Texas Commission on Environmental Quality (TCEQ). Understanding factors that cause Arsenic water pollution is of great interest to hydrologists.

Currently the GWDB has water quality data for 105,814 wells in Texas that have been collected over last 25 years. The database has to be cleaned of duplicate, missing and/or inconsistent values. As we are particularly interested in Arsenic, we have considered only those wells where there is at least one sample for Arsenic concentration. When multiple samples exist for a well, we take the average value. For each non-spatial attribute, we calculate z-scores and then calculate z(A↑,o) and z(A↓,o) using formulas (2) and (3). The particular dataset we used in the evaluation has 3 spatial attributes: longitude, latitude and aquifer, and 10 non-spatial attributes: Arsenic (As), Molybdenum (Mo), Vanadium (V), Boron (B), Fluoride (F^-), Silica (SiO_2), Chloride (Cl^-) and Sulfate (SO_4^{2-}) to which Total Dissolved Solids (TDS) and Well Depth (WD) are added. Those particular elements are chosen among the number of chemical elements available because of similar geochemical behavior—that is, travel together—(Mo, V) [24], or because those parameters could point out mobilizing mechanisms (Cl^-, SO_4^{2-}, TDS, well depth), or because they could suggest the ultimate origin of Arsenic (F^-, B, SiO_2). The created dataset contains average values of the 10 non-spatial attributes among 1,653 wells and no null values. Here onwards we call this dataset Arsenic_10_avg. We also created other datasets from GWDB that are available on the web [8].

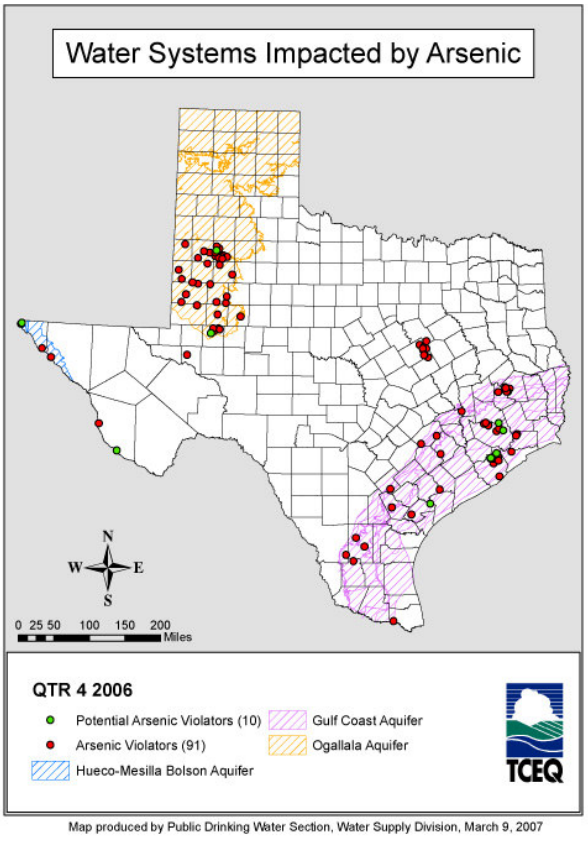

Figure 3. Arsenic pollution map (source TCEQ).

Table 2. Fitness function parameters used.

All experiments: P(B) = (As↑∈ B or As↓∈ B) and |B|<5; th=0, η=1.	
Experiment 1	β = 1.3, θ=1.0
Experiment 2	β = 1.5, θ=1.0
Experiment 3	β = 2.0, θ=1.0
Experiment 4	β = 1.5, θ=5.0

Table 3. Top 5 regions ranked by interestingness (as per formula 7).

Exp. No.	Top 5 Regions	Region Size	i(c)	Maximum Valued Pattern in the Region	Purity	Average Product for maximum valued pattern
Exp. 1	1	23	174.3191	As↑Mo↑V↑F^-↑	0.83	211.0179
	2	40	104.8576	As↑Mo↑V↑	0.65	161.3194
	3	11	92.9385	As↑Mo↑V↑SO_4^{2-}↑	0.55	170.3873
	4	36	89.4068	As↑B↑Cl^-↑TDS↑	0.58	153.2687
	5	7	30.5775	As↑Mo↑Cl^-↑TDS↑	0.57	53.5107
Exp. 2	1	80	33.5978	As↑B↑Cl^-↑TDS↑	0.48	70.7322
	2	181	25.3314	As↑Mo↑V↑F^-↑	0.49	52.1020
	3	17	6.4819	As↑Mo↑Cl^-↑TDS↑	0.29	22.0383
	4	23	6.4819	As↓Cl^-↑SO_4^{2-}↑TDS↑	0.78	8.1287
	5	10	3.4645	As↓B↑Cl^-↑TDS↑	0.4	8.6612
Exp. 3	1	238	5.3234	As↑B↑Cl^-↑TDS↑	0.22	23.9052
	2	833	1.8118	As↑Mo↑V↑F^-↑	0.16	11.4334
	3	152	0.3201	As↓SiO_2↑WD↑	0.53	0.6006
	4	432	0.1969	As↓TDS↓	0.93	0.2122
Exp. 4	1	7	630.1098	As↑B↑Cl^-↑TDS↑	1.0	630.1097
	2	2	541.4630	As↑Mo↑V↑B↑	1.0	541.4630
	3	1	466.8389	As↑B↑ SO_4^{2-}↑TDS↑	1.0	466.8389
	4	4	275.4066	As↑V↑ SO_4^{2-}↑TDS↑	1.0	275.4066
	5	3	234.7918	As↑Mo↑B↑SO_4^{2-}↑	1.0	234.7918

Experimental Results. We have tested our regional co-location mining framework by applying the algorithm CLEVER using the fitness function described in Section 3 to the above dataset. Because we are interested in discovering co-location patterns with respect to Arsenic, only co-location sets that contain As↑ or contain As↓ are considered. Table 2 summarizes the fitness function parameters used in the experiments. As the value of parameter β affects region size, we have conducted experiments using three different values for this parameter (Experiments 1-3); maximum co-location set size is restricted to four in these experiments. The parameter θ determines importance of purity when evaluating regions. To examine the impact of parameter θ, we design Experiment 4 using a very high θ value but otherwise the parameters are identical to those of Experiment 2. When using our methodology, we observed that domain experts are interested in both top ranked regions with respect to interestingness and reward. While ranking using interestingness highlights local outliers, ranking using reward identifies larger regions with more general patterns. Table 3 gives details of top 5 regions ranked by interestingness, and Table 4 visualizes these regions on the map of Texas. Table 5 describes the top 5 co-location regions ranked by reward, and Table 6 visualizes the top reward regions of experiments 2 and 4.

The parameter β affects the size of the co-location regions discovered. As illustrated in Table 4, as β increases from Experiment 1 to Experiment 3, CLEVER finds fewer, larger regions. For example, for β=2.0, CLEVER finds only 4 quite large regions capturing almost global patterns. The algorithm is able to determine known areas of high Arsenic concentrations as well as interesting unknown features. High Arsenic is a well-known problem in the Southern Ogallala Aquifer in the Texas Panhandle (rectangular area in northern Texas) and in the Southern Gulf Coast Aquifer north of the Mexican border (see Figure 3 for the aquifer locations). Figure 4 (Experiment 1) does recognize the higher Arsenic concentration areas in the Panhandle (ranks 1, 2, and 3 in color red, orange and yellow) associated with high Molybdenum and Vanadium, and is also able to discriminate among companion elements such as Fluoride (rank 1 region in red) or Sulfate (rank 3 region in yellow). The Gulf Coast area (rank 4 region in green) is characterized by a Boron marker, not present in the Panhandle, suggesting different Arsenic mobilization mechanisms. When the clusters are not as tightly defined (Figure 5 Experiment 2, larger β), they display the usually recognized extend of Arsenic contamination in Texas: Ogallala Aquifer, Southern Gulf Coast, and West Texas basins. Areas delimited by clusters of ranks 4 and 5 (illustrated in green and blue, respectively) are characterized by low Arsenic but general chemistry similar to the

high Arsenic cluster (rank 1 in red). A further loosening of cluster definition (Figure 6 Experiment 3) results in a display of the known, often described as sharp, boundaries between high and low Arsenic areas in the Ogallala Aquifer (ranks 2 and 4 in orange and green, respectively) and the Gulf Coast aquifer (ranks 1 and 3 in red and yellow, respectively). In addition, analysis of the Arsenic pollution map in Figure 3 and the algorithm results in Tables 3 and 4 clearly shows that our approach successfully identified all known regions with high Arsenic contamination.

The results also identify some inconsistent co-location sets in Table 3, Exp. 2 (Figure 5): the rank 3 region, depicted in yellow, located in the area of the Hueco-Mesilla Bolson Aquifer is characterized by the co-location set {As↑Mo↑Cl^-↑TDS↑} and the rank 5 region, depicted in blue, in the Gulf Coast Aquifer has co-location set {As↓B↑Cl^-↑TDS↑}: As↑ is co-located with Cl^-↑ and TDS↑ in one region but As↓ is co-located with Cl^-↑ and TDS↑ in the other region. As displayed in Figure 5, the rank 3 region (in yellow) is in West Texas, whereas the rank 5 region (in blue) is in South Texas. Our regional co-location mining framework successfully identifies such inconsistent regional patterns. The inconsistent patterns are not a problem as they are regional and not global patterns.

Moreover, as we increase the value of θ to 5, as expected, only co-location sets with purities above 90% are discovered. We also observe that, the region of {As↑Mo↑V↑F^-↑}, the maximum reward regions of Exp. 2 (Figure 8 in color red) and the region of {As↑V↑F^-↑}, the second ranked reward region of Exp. 4 (Figure 9 in color orange) occupy a similar spatial extent in North-West Texas. The first region is characterized by the co-location set {As↑Mo↑V↑F^-↑}, whereas the second region has the co-location set {As↑V↑F^-↑} associated with it and is slightly wider but significantly shorter than the first region. The dropping of Mo↑ from the co-location set increases purity 49% to 91%, but the average product drops from 52.1 to 12.8; this explains why the smaller co-location set is selected when θ is 5—but the larger set is better when θ is 1. When θ is decreased to 0, surprisingly, the complete dataset is returned as a single region with the co-location set of {As↑Mo↑V↑F^-↑} with an average product of 5.95 and a purity of only 0.086. Also, in the rank 3 reward region of the Exp. 2, the pattern {As↓TDS↓} is observed (Figure 8 in yellow); the interestingness of the pattern is quite low, but its purity is 0.91 and the region contains 467 wells. In general, when ranking regions by reward, region size becomes more important due to the additive nature of the employed fitness function.

Table 4. Top 5 regions ranked by interestingness: red=1st, orange=2nd, yellow=3rd, green=4th, blue=5th.

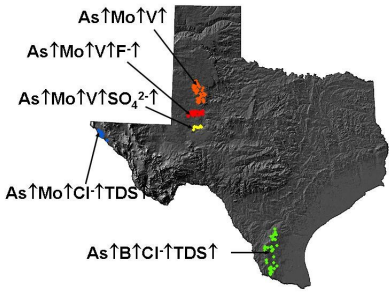

Figure 4 Experiment 1

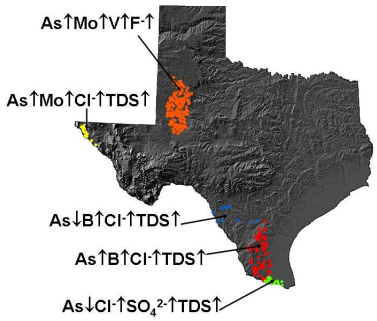

Figure 5 Experiment 2

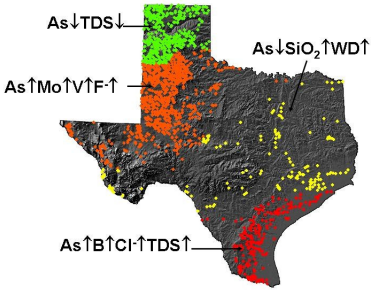

Figure 6 Experiment 3

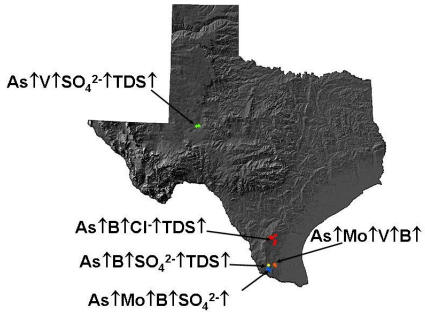

Figure 7 Experiment 4

Table 5. Top 5 regions ranked by reward (as per formula 8).

Exp. No.	Top 5 Regi-ons	Region Size	Region Reward	Maximum Valued Pattern in theRegion	Purity	Average Product for maximum valued pattern
Exp. 1	1	40	12684.6304	As↑Mo↑V↑	0.65	161.3194
	2	23	10270.49	As↑Mo↑V↑F^-↑	0.83	211.0179
	3	36	9431.1264	As↑B↑Cl^-↑TDS↑	0.58	153.2687
	4	11	2098.970187	As↑Mo↑V↑SO_4^{2-}↑	0.55	170.3873
	5	507	578.8116	As↓TDS↓	0.90	0.1968
Exp. 2	1	181	61684.5323	As↑Mo↑V↑F^-↑	0.49	52.1019
	2	80	24040.6315	As↑B↑Cl^-↑TDS↑	0.48	70.7322
	3	467	1884.8856	As↓TDS↓	0.91	0.2047
	4	23	701.7072	As↓Cl^-↑SO_4^{2-}↑TDS↑	0.78	8.1287
	5	189	587.9790	As↓F^-↓	0.78	0.2909
Exp. 3	1	833	1257170.945	As↑Mo↑V↑F^-↑	0.16	11.4334
	2	238	301539.908	As↑B↑Cl^-↑TDS↑	0.22	23.9052
	3	432	36754.1035	As↓TDS↓	0.93	0.2122
	4	152	7394.7640	As↓SiO_2↑WD↑	0.53	0.6006
Exp. 4	1	7	11669.7965	As↑B↑Cl^-↑TDS↑	1.0	630.1097
	2	117	10407.3250	As↑V↑F^-↑	0.91	12.8550
	3	4	2203.2526	As↑V↑ SO_4^{2-}↑TDS↑	1.0	275.4066
	4	2	1531.4887	As↑Mo↑V↑B↑	1.0	541.4630
	5	530	1426.9140	As↓TDS↓	0.90	0.1939

Table 6. Top 5 regions ranked by reward. Colors: red – 1st-ranked, orange – 2nd-ranked, yellow – 3rd-ranked, green – 4th-ranked, blue – 5th-ranked.

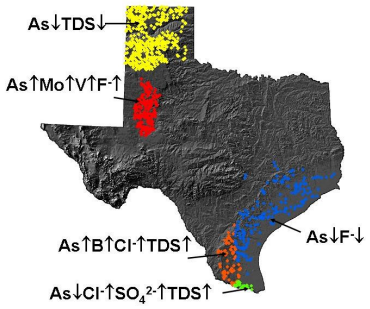

Figure 8 Experiment 2

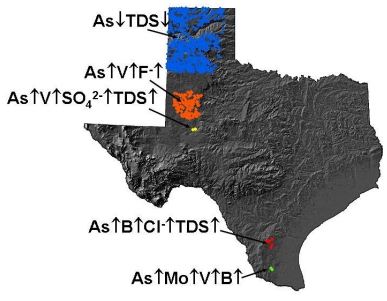

Figure 9 Experiment 4

Parameters and Multi-Run Analysis. When the co-location mining framework is used in the case study a lot of parameters have to be selected prior to running the mining algorithm. These parameters can be subdivided into fitness functions parameters and region discovery algorithm parameter. As far as fitness function parameters are concerned, they are selected in close collaboration with domain experts and a selection is made based on what kind of patterns the domain expert is interested in**.** However, it should be noted that domain experts are usually interested in obtaining regions and their associated patterns taking multiple perspectives which necessitates running region discovery algorithms multiple times with different fitness function parameter settings. For example, in the particular application the domain expert is interested in finding very small, local regions containing few objects with very large z-value products that more resemble outliers for further scientific investigation. Moreover, results that characterize associations between Arsenic (or the absence of Arsenic) and other chemicals at the regional level are also desirable. Consequently, because it is important to identify regions at different levels of granularity it is necessary to conduct experiment for multiple values of parameter β. In summary, for analyzing co-location relationships in the Arsenic data it is necessary to run the region discovery algorithm multiple times; e.g. for 30 different fitness function parameter settings. This raises the question what can be done to facilitate the analysis of results from multiple runs. To address this problem, we are currently developing a multi-run analysis system that stores the obtained regions and their associated properties in a spatial database. Regions themselves can be represented as polygons[5] which makes it easy to query them and to analyze relationships with respect to results of multiple runs automatically. For example, overlap between two regions can be computed as the size of the intersection of two region polygons.

As far as region discovery algorithm parameters—CLEVER in our case—are concerned, it is desirable to compute those automatically. In the particular experiment neighborhood sizes are chosen based on results of a complex experiment that compared different neighborhood sizes with respect to solution quality and overhead, and a neighborhood size of 3 is chosen which remained fixed for all experiments. However, other parameters, namely k', p, and p' where chosen automatically based on results of short runs that stopped after 20 iterations. Basically, the results of short runs are used to determine the utility of different parameter settings for the three parameters relying on a simple reinforcement procedure and CLEVER is then run with the "best" parameter setting. Run time used and quality of solutions found (measured by q(X)) provided the environmental feedback for the reinforcement learning procedure; basically, q(X) is maximized but algorithm runtime had to be bounded.

Performance. We also analyze the run-time needed to conduct the experiments. Our algorithms have been developed using an open source, Java-based data mining and machine learning framework *Cougar^2*, which is developed by our research group [7]. All the experiments are conducted on a machine with 1.3 GHz of processor speed and 4 GB of memory. The machine runs RedHat Enterprise Linux 3 on ia64 architecture. Our analysis shows that the CLEVER algorithm allocated more than 98% of its resources to the following two tasks: creating clusters for a given set of representatives and for fitness computations. With maximum pattern length set to 3, around 76% of time is allocated to computing q(X) and it takes around 1-2 hours for the algorithm to terminate. With maximum pattern length set to 4, 90% of the runtime is allocated to fitness computations and in most cases the algorithm terminates in 6-15 hours.

[5] Basically, CLEVER computes spatial clusters that are Voronoi cells in the 2D longitude-latitude space.

5. SUMMARY AND DISCUSSION

This paper proposes a novel framework for mining co-locations patterns in spatial datasets. In contrast to past co-location mining research that centers on finding global co-location patterns in categorical datasets, a regional co-location mining framework is introduced that operate in the continuous domain without need for discretization. The framework views regional co-location mining as a clustering problem in which an externally given reward-based fitness has to be maximized; in particular, fitness functions we employ in our approach, rely on products of z-scores of continuous variables to assess the interestingness of co-location patterns in the continuous space. A highly desirable feature of our approach is that it provides search-engine-like capabilities to scientists by returning regions ranked by the scientist's notion of interestingness that has been captured in a plug-in fitness function.

The framework is evaluated in a case study involving chemical concentrations of Texas water wells centering on co-location patterns involving Arsenic. The tested region discovery algorithm is able to identify known and unknown regional co-location sets. Different sets of algorithm parameters lead to the characterization of Arsenic distribution at different levels of granularity—stressing the need for parameterized, plug-in fitness functions that allow domain experts to express what patterns they are looking for at what level a granularity.

Arsenic water pollution is a serious problem for Texas and its causes are complex and frequently difficult to explain, particularly for wells in the Ogallala aquifer [22]. A large number of possible explanations exist what causes high levels of Arsenic concentrations to occur. Therefore, scientists face the problem to decide which promising hypotheses from a large set of hypotheses to be investigated further. The proposed framework is particularly useful in the early stages of a research study when domain scientists are exposed to massive amount of data with only a few clues to organize them. In general, our regional co-location mining framework turned out to be valuable to domain experts in that it provides a data driven approach which suggests promising hypotheses for future research. In particular, unexpected associations selected by the framework can challenge preconceived ideas and open the way to potential breakthroughs in the study of Arsenic subsurface contamination.

Finally, a novel, prototype-based region discovery algorithm named CLEVER has been introduced that seeks the optimal number of clusters, uses larger neighborhood sizes to battle premature convergence, and uses randomized hill climbing and re-sampling to reduce algorithm complexity.

REFERENCES

[1] Achtert, E., Böhm, C., Kriegel, H., Kröger, P., and Zimek, A. 2006. Deriving Quantitative Models for Correlation Clusters. In Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining (Philadelphia, PA, USA, August 2006). KDD '06. ACM, New York, NY, 4-13.

[2] Aggarwal, C. C., Procopiuc,C. M., and Yu, P. S. 2002. Finding Localized Associations in Market Basket Data. IEEE Transactions on Knowledge and Data Engineering, 14, 51-62.

[3] Aumann, Y., and Lindell, Y. 1999. A Statistical Theory For Quantitative Association Rules. In Proc. of the 5th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. KDD '99. ACM, New York, NY, 261-270.

[4] Brimicombe, A. J. 2005. Cluster Detection in Point Event Data Having Tendency Towards Spatially Repetitive Events. In the 8th Intl. Conf. on GeoComputation.

[5] Calders, T., Goethals, B., and Jaroszewicz, S. 2006. Mining Rank-Correlated Sets of Numerical Attributes. In Proc. of the 12th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining. KDD '06. ACM, New York, NY, 96-105.

[6] Choo, J., Jiamthapthaksin, R., Chen, C., Celepcikay, O., Giusti, C., and Eick, C. F. 2007. MOSAIC: A Proximity Graph Approach to Agglomerative Clustering. In Proc. of the 9th Intl. Conf. on Data Warehousing and Knowledge Discovery. DaWaK' 07.

[7] Cougar^2: Data Mining and Machine Learning Framework, https://cougarsquared.dev.java.net/.

[8] Data Mining and Machine Learning Group, University of Houston, http://www.tlc2.uh.edu/dmmlg.

[9] Ding, W., Eick, C. F., Wang, J., and Yuan, X. 2006. A Framework for Regional Association Rule Mining in Spatial Datasets. 2006. In Proc. of the IEEE Intl. Conf. on Data Mining. ICDM'06.

[10] Ding, W., Jiamthapthaksin, R., Parmar, R., Jiang, D., Stepinski, T., and Eick, C. F. 2008. Towards Region Discovery in Spatial Datasets. In Proc. of Pacific-Asia Conference on Knowledge Discovery and Data Mining (Osaka, Japan, May 2008). PAKDD '08.

[11] Eick, C.F., Vaezian, B., Jiang, D., and Wang, J. 2006. Discovering of Interesting Regions in Spatial Data Sets Using Supervised Clustering. In Proc. of the 10th European Conference on Principles of Data Mining and Knowledge Discovery. PKDD '06.

[12] Eick, C. F., Zeidat, N., and Zhao, Z. Supervised Clustering --- Algorithms and Benefits. In Proc. of the Intl. Conf. on Tools with AI (Boca Raton, Florida, November 2004). ICTAI '04, 774-776.

[13] Getis, A., and Ord, J. K. 1996. Local Spatial Statistics: an Overview. In Spatial analysis: modeling in a GIS environment, Cambridge, GeoInformation International. (Cambridge, 1996), 261–277.

[14] Huang, Y., Pei, J., and Xiong, H. 2006. Mining Co-Location Patterns with Rare Events from Spatial Data Sets. Geoinformatica 10 (3), 239-260.

[15] Huang, Y., and Zhang, P. 2006. On the Relationships between Clustering and Spatial Co-location Pattern Mining. In Proc. of the 18th IEEE Intl. Conf. on Tools with Artificial intelligence. ICTAI. IEEE Computer Society, Washington, DC, 513-522.

[16] Jaroszewicz, S. 2008. Minimum Variance Associations—Discovering Relationships in Numerical Data. In Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, (Osaka, Japan, May 2008). PAKDD '08.

[17] Kaufman, L., and Rousseeuw, P. J. 2005. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New Jersey.

[18] Kulldorff, M. 2001. Prospective Time Periodic Geographical Disease Surveillance Using a Scan Statistic. Journal of the Royal Statistical Society Series A, 164, 6-72.

[19] Lloyd, S.P. 1982. Least Squares Quantization in PCM. IEEE Trans. on Information Theory, 28, 128–137.

[20] Ng, R. T., and Han, J. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining. In Proc. of the 20th Intl. Conf. on Very Large Data Bases. Morgan Kaufmann Publishers, San Francisco, CA, 144-155.

[21] Ord, J. K., and Getis, 1995. A. Local Spatial Autocorrelation Statistics: Distributional Issues and an Application. Geographical Analysis, 27(4), 286-306.

[22] Scanlon, B. R., Nicot, J. P. et al. 2005. Evaluation of Arsenic Contamination in Texas. Technical report prepared for TCEQ, under contract no. UT-08-5-70828.

[23] Shekhar, S., and Huang, Y. 2001. Discovering Spatial Co-location Patterns: A Summary of Results. In Proc. of the 7th Intl. Symp. on Advances in Spatial and Temporal Databases, Springer-Verlag, London, 236-256.

[24] Smedley, P. L., and Kinniburgh, D. G. 2002. A Review of the Source, Behavior and Distribution of Arsenic in Natural Waters. Applied Geochemistry 17, 517-568.

[25] Smith, A. H. et al. 1992. Cancer Risks From Arsenic in Drinking Water. Environmental Health Perspectives, 97, 259–267.

[26] Srikant, R., and Agrawal, R. 1996. Mining Quantitative Association Rules in Large Relational Tables. SIGMOD Rec. 25(2), 1-12.

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

[28] Xiong, H., Shekhar, S., Huang, Y., Kumar, V., Ma, X., and Yoo, J. S. 2004. A Framework for Discovering Co-location Patterns in Data Sets with Extended Spatial Objects. In Proc. Of SIAM Intl. Conf. on Data Mining (SDM).

[29] Yoo, J.S., and Shekhar, S. 2006. A Join-less Approach for Mining Spatial Co-location Patterns. IEEE Transactions on Knowledge and Data Engineering (TKDE), 18.