

# Regional Pattern Discovery in Geo-Referenced Datasets Using PCA

Oner Ulvi Celepcikay<sup>1</sup>, Christoph F. Eick<sup>1</sup>, and Carlos Ordonez<sup>1</sup>,

<sup>1</sup> University of Houston, Department of Computer Science, Houston, TX, 77204-3010  
(*onerulvi, ceick, ordonez*)@cs.uh.edu

**Abstract.** Existing data mining techniques mostly focus on finding global patterns and lack the ability to systematically discover regional patterns. Most relationships in spatial datasets are regional; therefore there is a great need to extract regional knowledge from spatial datasets. This paper proposes a novel framework to discover interesting regions characterized by “strong regional correlation relationships” between attributes, and methods to analyze differences and similarities between regions. The framework employs a two-phase approach: it first discovers regions by employing clustering algorithms that maximize a PCA-based fitness function and then applies post processing techniques to explain underlying regional structures and correlation patterns. Additionally, a new similarity measure that assesses the structural similarity of regions based on correlation sets is introduced. We evaluate our framework in a case study which centers on finding correlations between arsenic pollution and other factors in water wells and demonstrate that our framework effectively identifies regional correlation patterns.

**Keywords:** Spatial Data Mining, Correlation Patterns, Regional Knowledge Discovery, Clustering, PCA.

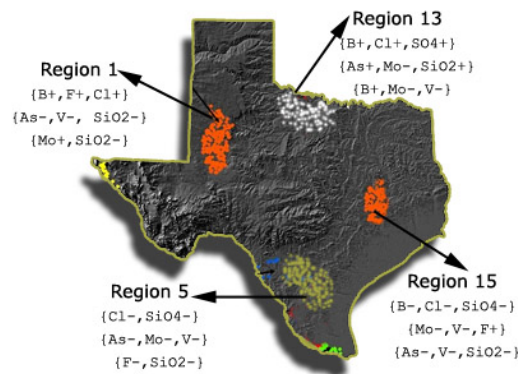
## 1 Introduction

Advances in database and data acquisition technologies have resulted in an immense amount of geo-referenced data, much of which cannot be adequately explored using current methodologies. The goal of spatial data mining is to automate the extraction of interesting and useful patterns that are not explicitly represented in geo-referenced datasets. Of particular interest to scientists are techniques which are capable of finding scientifically meaningful regions and representing their associated patterns in spatial datasets, as such techniques have many immediate applications in medicine, geosciences, and environmental sciences, such as the association of particular cancers with environmental pollution of sub-regions, the detection of crime zones with unusual activities, and the identification of earthquake hotspots. Since most relationships in spatial datasets are geographically regional [15], there is a great need to discover regional knowledge in spatial datasets. Existing spatial data mining techniques mostly focus on finding global patterns and lack the ability to systematically discover regional patterns. For example, a strong correlation between a fatal disease and a set of chemical concentrations in water wells might not be

detectable throughout Texas, but such a correlation pattern might exist regionally which is also a reflection of Simpsons' paradox[16]. This type of regional knowledge is crucial for domain experts who seek to understand the causes of such diseases and predict future cases. Another issue is that regional patterns have a scope that—because they are not global—is a subspace of the spatial space. This fact complicates their discovery because both subspaces and patterns have to be searched. Work by Celik et al. [4] assumes the presence of an apriori given regional structure (e.g. a grid) and then searches for regional patterns. One unique characteristic of the framework presented in this paper is that it searches for interesting subspaces by maximizing a plug-in reward-based interestingness function and then extracts regional knowledge from the obtained subspaces.

This paper focuses on discovering regional correlation patterns that are associated with contiguous areas in the spatial subspaces, which we call regions. Interesting regions are identified by running a clustering algorithm that maximizes a PCA-based fitness function. PCA is used to guide the search for regions with strong structural relationships. Figure 1 shows an example of discovered regions along with their highest correlated attribute sets (HCAS). For example, in Region 1 a positive correlation between Boron (B), Fluoride (F), and Chloride (Cl), and between Arsenic (As), Vanadium (V), and Silica (SiO<sub>2</sub>), as well as a negative correlation between Silica (SiO<sub>2</sub>) and Molybdenum (M) can be observed. As can be seen in the Figure 1, some of those sets differ quite significantly between regions, emphasizing the need for regional knowledge discovery.

Also a new similarity measure is introduced to estimate the structural similarity between regions based on correlation sets that are associated with particular regions. This measure is generic and can be used in other contexts when two sets of principal components have to be compared.



**Fig. 1.** An Example of Regional Correlation Patterns for Chemical Concentrations in Texas

The main contributions of the paper are:

1. A framework to discover interesting regions and their regional correlation patterns.
2. A PCA-based fitness function to guide the search for regions with well-defined PCs
3. A generic similarity measure to assess the similarity between regions quantitatively.
4. An experimental evaluation of the framework in a case study that centers on indentifying causes of arsenic contamination in Texas water wells.

The remainder of the paper is organized as follows: In section 2, we discuss related work. In section 3, we provide a detailed discussion of our region discovery framework, the PCA-based fitness function and HCAS similarity measure. Section 4 presents the experimental evaluation and section 5 concludes the paper.

## 2 Related Work

**Principal Component Analysis (PCA):** PCA is a multivariate statistical analysis method that is very commonly used to discover highly correlated attributes and to reduce dimensionality. The idea is to identify  $k$  principal components for an  $d$ -dimensional dataset ( $k \ll d$ ) that explain a large portion of the dataset's variance, e.g. more than 80%, which allows the reduction of the dataset's dimensionality from  $d$  to  $k$  dimension without much information loss. PCA is widely used for data mining and some PCA-based clustering methods have been developed in the past [13, 14]. PCA has also been extensively applied extensively in the field of face recognition [18]. The authors in [20] proposed a supervised PCA model called SPPCA, whereby they extended PCA to incorporate label/class information into the projection phase.

**Correlation Clustering:** Correlation clustering aims at grouping the data into correlation clusters such that the objects in the same cluster exhibit a certain density and objects are all associated with the same arbitrarily oriented hyperplane of arbitrary dimensionality [1]. The 4C algorithm [3] combines PCA and DBSCAN in order to identify correlation connected clusters that are subgroups of data points exhibiting similar correlations. One drawback of the 4C algorithm is that the user must choose proper values for many algorithm parameters. Correlation clustering is similar to our work in that it deals with the identification of correlation clusters; however, our framework uses an external PCA-based plug-in fitness function for clustering, and it is applicable in conjunction with any clustering algorithm that supports plug-in fitness functions, whereas 4C is dependent on the DBSCAN clustering framework. For example, our framework could be used in conjunction with the agglomerative clustering algorithm MOSAIC [5] which discovers arbitrary shaped regions. Moreover, correlation clusters are not necessarily contiguous in the attribute space, whereas our approach identifies contiguous regions in the spatial subspace that display strong principal components with respect to the non-spatial attributes.

## 3 Methodology

We now present the methods our regional pattern discovery framework utilizes during the region discovery and post processing phases as illustrated in Figure 2.

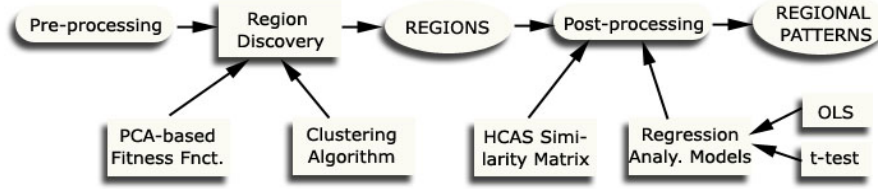


Fig. 2. Regional Pattern Discovery Framework

### 3.1. Region Discovery Framework

We employ the region discovery framework that was proposed in [10, 11]. The objective of region discovery is to find interesting places in spatial datasets—regions occupying contiguous areas in the spatial subspace. In this work, we extend this framework to find regional correlation patterns. The framework employs a reward-based evaluation scheme to evaluate the quality of the discovered regions. Given a set of regions  $R = \{r_1, \dots, r_k\}$  with respect to a spatial dataset  $O = \{o_1, \dots, o_n\}$ , the fitness of  $R$  is defined as the sum of the rewards obtained from each region  $r_j$  ( $j = 1, \dots, k$ ):

$$q(R) = \sum_{j=1}^k i(r_j) * size(r_j)^\beta . \quad (1)$$

where  $i(c_j)$  is the interestingness of the region  $r_j$ —a quantity based on domain interest to reflect the degree to which the region is “newsworthy”. The framework seeks for a set of regions  $R$  such that the sum of rewards over all of its constituent regions is maximized. In general, the parameter  $\beta$  controls how much premium is put on region size. The  $size(r_j)^\beta$  component in  $q(R)$ , ( $\beta \geq 1$ ) increases the value of the fitness nonlinearly with respect to the number of objects in 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. Rewarding region size non-linearly ensures merging neighboring regions whose PCs are structurally similar.

**The CLEVER Algorithm:** We employ the CLEVER [10] clustering algorithm to find interesting regions in the experimental evaluation. CLEVER is a representative-based clustering algorithm that forms clusters by assigning objects to the closest cluster representative. The algorithm starts with a randomly created set of representatives and employs randomized hill climbing by sampling  $s$  neighbors of the current clustering solution as long as new clustering solutions improve the fitness value. To battle premature convergence, the algorithm employs re-sampling: if none of the  $s$  neighbors improves the fitness value, then  $t$  more solutions are sampled before the algorithm terminates. In short, CLEVER searches for the optimal set of regions, maximizing a given, plug-in fitness function  $q(R)$ , which in our case is the PCA-based fitness function.

### 3.2. PCA-based Fitness Function for Region Discovery

The directions identified by PCA are the eigenvectors of the correlation matrix. Each eigenvector has an associated eigenvalue that is a measure of the corresponding variance and the PCs are ordered with respect to the variance associated with that component in descending order. Ideally, it is desirable to have high eigenvalues for the first  $k$  PCs, since this means that a smaller number of PCs will be adequate to account for the threshold variance which overall suggests that a strong correlation among variables exists[14]. Our work employs the interestingness measure in definition 1 to assess the strength of relationships between attributes in a region  $r$ :

**Definition 1:** (PCA-based Interestingness –  $i_{PCA}(r)$ )

Let  $\lambda_1, \lambda_2, \dots, \lambda_k$  be the eigenvalues of the first  $k$  PCs, with  $k$  being a parameter:

$$i_{PCA}(r) = \sqrt{(\lambda_1^2 + \dots + \lambda_k^2)/k} . \quad (2)$$

PCA-based fitness function then becomes:

$$q_{PCA}(R) = \sum_{j=1}^k i_{PCA}(r_j) * size(r_j)^\beta . \quad (3)$$

The fitness function rewards high eigenvalues for the first  $k$  PCs. By taking the square of each eigenvalue we ensure that regions with a higher spread in their eigenvalues will obtain higher rewards—reflecting the higher importance assigned in PCA to higher ranked principal components. For example; a region with eigenvalues  $\{6, 2, 1, \dots\}$  will get a higher reward than a region with eigenvalues  $\{4, 3, 2, \dots\}$  even though the total variance captured in both cases is about the same.

We developed a generic pre-processing technique to select the best  $k$  value for the PCA-based fitness function for a given dataset that is based on a variance threshold: the smallest  $k$  is chosen so that the variance captured in the first  $k$  principal components is greater than this threshold. First, the algorithm applies PCA to the global data and determines the global  $k$  value ( $k_g$ ) for a given variance threshold which serves as an upper bound for  $k$ . Then, it splits the spatial data into grids (random square regions), applies PCA to each grid, and determines the  $k$  value for each region based on the variance threshold obtaining  $\{k_{r1}, \dots, k_{rs}\}$ . The algorithm next selects the most frequent  $k_r$  value in the set of regional  $k$ -values as the final result—to be used in the fitness function. For datasets with strong regional patterns, the chosen  $k$  is expected to be lower than  $k_g$ : fewer PCs capture the same variance in the regional data, because regional correlation is stronger than global correlation.

Our fitness function repeatedly applies PCA during the search for the optimal set of regions, maximizing the eigenvalues of the first  $k$  PCs in that region. Having an externally plugged in PCA-based fitness function enables the clustering algorithm to probe for the optimal partitioning and encourages the merging of two regions that exhibit structural similarities. This approach is also more advantageous than applying PCA once or multiple times on the data, since the PCA-based fitness function is applied repeatedly to candidate regions to explore each possible region combination.

### 3.3. Correlation Sets, HCAS and Region Similarity

Highest correlated attribute sets (HCAS) are sets of correlation sets (CSs) which are signed sets of the attributes that are highly correlated. CSs are constructed from the eigenvectors of principal components (PCs). An attribute is added to the correlation set of a PC, if the absolute value of the PC coefficient of that attribute is above a threshold  $\alpha$  along with the sign of the coefficient. The threshold  $\alpha$  is selected based on the input from domain experts. For example, let's assume that  $\alpha=0.33$  for  $PC_1$  in Table 1, a correlation set  $\{Mo-, Cl+, SiO4+\}$  is constructed, since only the absolute values of these attributes' coefficients are above  $\alpha$  (depicted in bold in the table). In this set, Mo is negatively correlated with both Cl and SiO4, whereas Cl and SiO4 are positively correlated.

**Table 1.** Eigen-Vectors of first k PCs (k=3)

| Variables        | PC <sub>1</sub> | PC <sub>2</sub> | PC <sub>3</sub> |
|------------------|-----------------|-----------------|-----------------|
| As               | -0.323          | <b>-0.452</b>   | <b>-0.34</b>    |
| Mo               | <b>-0.346</b>   | 0.062           | <b>0.46</b>     |
| V                | -0.301          | <b>-0.463</b>   | -0.243          |
| B                | 0.138           | -0.247          | -0.177          |
| F                | 0.325           | 0.177           | <b>0.356</b>    |
| SiO <sub>2</sub> | -0.165          | <b>-0.389</b>   | -0.299          |
| Cl               | <b>0.394</b>    | -0.295          | 0.148           |
| SiO <sub>4</sub> | <b>0.363</b>    | -0.277          | 0.194           |
| TDS              | -0.243          | -0.323          | 0.173           |
| WD               | 0.097           | -0.233          | -0.226          |

Next, CS and HCAS will be defined formally, and similarity measures for correlation sets and regions will be introduced.

**Definition 2:** (Correlation Sets – CS)

*A CS is a set of signed attributes that capture correlation patterns.*

**Definition 3:** (Highest Correlated Attribute Sets – HCAS)

*HCAS are sets of correlation sets and they are used to summarize correlation relationships of regions.*

Each region has a HCAS of cardinality k since our framework retains only k PCs. Each CS is associated with a single PC (principal component). HCAS are constructed for each region to summarize their regional correlation patterns. HCAS are used to describe and compare the correlation patterns of regions.

**Example:** For the PCA result in Table1 the following HCAS will be generated:

$$\{\{Mo-, Cl+, SiO4+\}, \{As-, V-, SiO2-\}, \{As- Mo+, F+\}\}$$

Next, we define operations to manipulate CS.

**Definition 4:** (Operations on CSs)

Two operations are defined on CSs. They are;

1. **csign** (“complement sign”) changes the signs of a CS.  
e.g.  $csign(\{ A+, B-, C- \}) = \{ A-, B+, C+ \}$
2. **uns** (“unsigned”) removes the signs of attributes in a CS.  
e.g.  $uns(\{ A+, B-, C- \}) = \{ A, B, C \}$

**Definition 5:** (Correlation Sets Similarity – *simCS*)

The similarity between two correlation sets,  $CS_i$  and  $CS_j$ , is estimated using following equation:

$$simCS(CS_i, CS_j) = \frac{\max\{|CS_i \cap CS_j|, |CS_i \cap csign(CS_j)|\}}{|uns(CS_i) \cup uns(CS_j)|} \quad (4)$$

$simCS(CS_i, CS_j)$  is assessed by comparing  $CS_i$  with  $CS_j$  and comparing  $CS_i$  with  $csign(CS_j)$  and by taking the maximum set size obtained for the two intersections and dividing it by the number of objects in the union of unsigned  $CS_i$  and  $CS_j$ . Basically,  $simCS(CS_i, CS_j)$  takes two factors into consideration when comparing two CSs:

1. Agreement with respect to attributes that contribute to variance.
2. Agreement in correlation with respect to common attributes.

**Examples:**

- a.  $simCS(\{A-, B-\}, \{A+, B+\}) = 1$
- b.  $simCS(\{A+, B+\}, \{A+, B-\}) = 0.5$
- c.  $simCS(\{A+, B+, C+, D+, E+\}, \{A+, B+, C-, D-, E-\}) = 0.6$

Next, we define a similarity measure to assess the similarity of two regions with respect to their  $k$  PCs. Let us assume that the HCASs of region  $R_1$  and region  $R_2$  are  $\{CS_1, \dots, CS_k\}$  and  $\{CS'_1, \dots, CS'_k\}$ , respectively, and that the principal components of the regions have eigenvalues  $\lambda_1, \dots, \lambda_k$  and  $\lambda'_1, \dots, \lambda'_k$ , respectively.

**Definition 6:** (PC Similarity Matrix – *PCS*)

Let *PCS* be a  $k \times k$  similarity matrix whose entries  $pcs(i, j)$  store the similarity between  $i^{th}$  correlation set of region  $R_1$  and  $j^{th}$  correlation set of region  $R_2$  weighted by the eigenvalues of the associated principal components.

$$PCS(i, j) = simCS(CS_i, CS'_j) * \delta_{ij} \quad (5)$$

where

$$\delta_{i,j} = \frac{\lambda_i + \lambda'_j}{\sum_{i=1}^k \lambda_i + \sum_{j=1}^k \lambda'_j}$$

We use  $\delta_{ij}$  to weigh in the contribution of a correlation set of a PC to the overall similarity based on its eigenvalue. If the eigenvalue is high, then its contribution when assessing similarity between regions is higher compared to other sets with lower eigenvalues. Next, using *PCS*, we introduce the regional similarity measure.

**Definition 7:** (Regional Similarity – *SimR*)

Let  $perm(k)$  be the set of all permutations of numbers  $1,..k$ , the similarity between two regions  $R_1$  and  $R_2$  is defined as follows:

$$simR(R_1, R_2) = \max_{y=(y_1, \dots, y_k) \in perm(k)} \sum_{i=1}^k PCS(i, y_i) \quad (6)$$

This similarity function computes an injective mapping from  $k$  principal components of region  $R_1$  to  $k$  principal components of region  $R_2$  which maximizes correlation set similarity weighted by the eigenvalues of the associated principal components. After the best injective mapping  $\psi$  has been determined, similarity is computed by adding the similarities of principal component  $i$  of region  $R_1$  with  $\psi(i)$  in region  $R_2$  for  $i=1, \dots, k$ . Basically,  $simR( )$  finds the best one-to-one mapping that aligns the principal components of the two regions to provide the best match with respect to the similarity. It should be noted that  $k$  is usually very small; typically 2-6, rarely larger than 10; therefore, maximizing similarity over all permutations usually can be done quickly. For larger  $k$  values, some greedy, approximate versions of the similarity function can be developed.

**Example:** Let us assume that the  $PCS$  of two regions ( $R_1$  and  $R_2$ ) is as follows: ( $k=3$  and  $S_1, S_2, S_3$  belongs to region  $R_1$  and  $T_1, T_2, T_3$  belongs to region  $R_2$ ):

| PCS            | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> |
|----------------|----------------|----------------|----------------|
| S <sub>1</sub> | 0.4            | 0.1            | 0.1            |
| S <sub>2</sub> | 0.0            | 0.1            | 0.3            |
| S <sub>3</sub> | 0.1            | 0.2            | 0.0            |

Since  $k$  is 3, there will be 6 one-to-one mappings between the principal components of the two regions. The similarity calculations that will be conducted to determine the similarity of  $R_1$  and  $R_2$  are shown below:

| R1             |                |                | Mappings | R2             |                |                | Calculations             |
|----------------|----------------|----------------|----------|----------------|----------------|----------------|--------------------------|
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>1</sub> | T <sub>2</sub> | T <sub>3</sub> | 0.4+0.1+0.0 = <b>0.5</b> |
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>1</sub> | T <sub>3</sub> | T <sub>2</sub> | 0.4+0.3+0.2 = <b>0.9</b> |
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>2</sub> | T <sub>1</sub> | T <sub>3</sub> | 0.1+0.0+0.0 = <b>0.1</b> |
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>2</sub> | T <sub>3</sub> | T <sub>1</sub> | 0.1+0.3+0.1 = <b>0.5</b> |
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>3</sub> | T <sub>1</sub> | T <sub>2</sub> | 0.1+0.1+0.2 = <b>0.4</b> |
| S <sub>1</sub> | S <sub>2</sub> | S <sub>3</sub> | --->     | T <sub>3</sub> | T <sub>2</sub> | T <sub>1</sub> | 0.1+0.1+0.1 = <b>0.3</b> |

In this case, the 2<sup>nd</sup> mapping  $\{S_1 \rightarrow T_1, S_2 \rightarrow T_3, S_3 \rightarrow T_2\}$  maximizes the sum of similarities.

So, we obtain  $SimR(R_1, R_2) = 0.9$



### 3.4. Post Processing via Regression Analysis

We additionally employ regression analysis models in the post processing phase to analyze regional dissimilarities. We use the OLS (Ordinary Least Squares) regression to investigate the impact of our independent variables on the dependent variable (e.g. arsenic concentration in arsenic experiments). OLS was chosen because it minimizes the mean squared error; thus, it is the best liner efficient estimator [19]. Our framework first applies regression analysis on global data (global regression); then, after it discovers regions, it retrieves the top k regions ranked by their interestingness and applies regression analysis on those regions (regional regression). The results of regional regression are compared with the results of global regression to reveal regional differences

## 4 Experimental Evaluation

### 4.1. A Real World Case Study: Texas Water Wells Arsenic Project

Arsenic is a deadly poison and even long-term exposure to very low arsenic concentration can cause cancer [17]. So it is extremely crucial to understand the factors that cause high arsenic concentrations to occur. In particular, we are interested in identifying other attributes that contribute significantly to the variance of arsenic concentration. Datasets used in the experiments were created using the Texas Water Department Ground Water Database [17] that samples Texas water wells regularly. The datasets were generated by cleaning out duplicate, missing and inconsistent variables and aggregating the arsenic amount when multiple samples exist. Our dataset has 3 spatial and 10 non-spatial attributes. Longitude, Latitude and Aquifer ID are the spatial attributes and Arsenic(As), Molybdenum(M), Vanadium(V), Boron(B), Fluoride(F), Silica(SiO<sub>2</sub>), Chloride(Cl), Sulfate(SiO<sub>4</sub>) are 8 of the non-spatial attributes which are chemical concentrations. The other 2 non-spatial attributes are Total Dissolved Solids (TDS) and Well Depth (WD). The dataset has 1,653 objects.

### 4.2. Experimental Parameters

Table 2 summarizes the common parameters used in all experiments and the ones specific to the individual experiments. These parameter values were chosen after many initial experiments as the parameter settings that provide the most interesting results.  $\beta$  is a parameter of the region discovery framework which controls the size of the regions to be discovered.  $s$  and  $t$  are the parameters of CLEVER algorithm. *min\_regon\_size* is a controlling parameter to battle the tendency towards having very small size regions with maximal variance. Regions with size below this parameter receive a reward of zero.

**Table 2.** The parameters used in the experiments

| Common parameters | $s=50, t=50, \alpha=0.33$              |
|-------------------|--|
| Experiment 1      | $min\_region\_size = 8, \beta = 1.7$   |
| Experiment 2      | $min\_region\_size = 9, \beta = 1.6$   |
| Experiment 3      | $min\_region\_size = 20, \beta = 1.7$  |
| Experiment 4      | $min\_region\_size = 16, \beta = 1.6$  |
| Experiment 5      | $min\_region\_size = 16, \beta = 1.01$ |

In our pre-processing phase to select the best  $k$  value for the experiment, we use 70% variance as the threshold, a percentage based on the comments from domain experts who maintain that this is a good threshold for detecting correlations among chemical concentrations. Other feedback from domain experts for the Water Pollution Experiment suggested that the arsenic dataset is not globally highly correlated; hence, setting the variance threshold to 70% is a good fit. The preprocessing phase, using 70% as the threshold, indicated that 3 or 4 are the best values for  $k$ . We report the results for  $k=3$  in this section. The threshold used in constructing correlation sets of HCAS was chosen in accordance with domain experts' feedback as 0.33.

### 4.3. HCAS and Similarity Results

HCASs for the experiment 1 are shown in Table 3 which lists the top 5 regions ranked by their interestingness values. These sets suggest that there are regional patterns involving highly correlated attributes, whereas globally (Texas-wide) almost all attributes are members of HCAS and are equally correlated; a situation which fails to reveal strong structural relationships. Analyzing the correlation sets and region similarity helps us to identify regions that display variations over space. For example, with respect to the second principal component of region 15 we observe a positive correlation between Molybdenum and Vanadium and a negative correlation between Molybdenum and Fluoride neither of which exists globally. Moreover, the negative correlation between Molybdenum and Fluoride only exist in region 15 and is not observed in the other four regions. In general, such observations are highly valuable to domain experts, because they identify interesting hypotheses and places for further investigation. Table 4 shows the regional similarity matrix and Table 5 depicts the similarity between the 5 regions and the global data (Texas).

**Table 3:** HCAS sets for the Top Ranked Regions

| Region ID    | HCAS Sets for the first 3 PCs |                            |                           |
|--------------|-------------------------------|----------------------------|---------------------------|
| <b>Texas</b> | $\{As-, Mo-, B-, Cl-, SO4-\}$ | $\{As+, V+, Fl+, SiO_2+\}$ | $\{As-, Mo-, SiO_2+\}$    |
| Region 0     | $\{Cl-, SO4-\}$               | $\{As-, Mo-, V-\}$         | $\{Fl-, SiO_2-\}$         |
| Region 1     | $\{B+, FL+, Cl+, SO4+\}$      | $\{As-, V-, SiO_2-\}$      | $\{Mo+, SiO_2-\}$         |
| Region 13    | $\{B+, Cl+, SO4+\}$           | $\{As+, Mo-, SiO_2+\}$     | $\{As-, Mo-, V-\}$        |
| Region 21    | $\{Mo+, B+, SiO_2-\}$         | $\{As-, V-, Cl+\}$         | $\{As+, Fl+, Cl+, SO4+\}$ |
| Region 15    | $\{B-, Cl-, SO4-\}$           | $\{Mo-, V-, Fl+\}$         | $\{As-, V-, SiO_2-\}$     |

**Table 4.** Similarity Matrix of Regions for Experiment 1

|           | Region 0 | Region 1 | Region 13   | Region 21 | Region 15   |
|-----------|----------|----------|-------------|-----------|-------------|
| Region 0  | 0.00     | 0.61     | 0.88        | 0.62      | 0.65        |
| Region 1  | 0.61     | 0.00     | 0.90        | 0.64      | 0.59        |
| Region 13 | 0.88     | 0.90     | 0.00        | 0.40      | <b>0.92</b> |
| Region 21 | 0.62     | 0.64     | 0.40        | 0.00      | 0.40        |
| Region 15 | 0.65     | 0.59     | <b>0.92</b> | 0.40      | 0.00        |

**Table 5.** Similarity Vector of Regions to Global Data (Texas) for Experiment 1

| Exp#1  | Region 0 | Region 1 | Region 13   | Region 21 | Region 15 |
|--------|----------|----------|-------------|-----------|-----------|
| Global | 0.59     | 0.70     | <b>0.84</b> | 0.59      | 0.66      |

**Discussion:** The HCAS similarity matrix and similarity vector in Tables 4 and 5 reveal that the HCAS similarity measure capture the true similarity between correlation patterns of regions. For example, we observe that region 13 and region 15 are the most similar regions. This is the result of following mappings:

$\{B+, Cl+, SO_4+\} \rightarrow \{B-, Cl-, SO_4-\}$  (  $PC_1$  of Region 13  $\rightarrow$   $PC_1$  of Region 15 mapping )  
 $\{As+, Mo-, SiO_2+\} \rightarrow \{As-, V-, SiO_2-\}$  (  $PC_2$  of Region 13  $\rightarrow$   $PC_3$  of Region 15 mapping )  
 $\{As-, Mo-, V-\} \rightarrow \{Mo-, V-, Fl+\}$  (  $PC_3$  of Region 13  $\rightarrow$   $PC_2$  of Region 15 mapping )

The discovered regions also maximize the cumulative variance captured through the first k principal components. The variance values for the top 5-ranked regions in experiment 2 are given in Table 6. The global data has a 57% cumulative variance which indicates that the attributes in the global data are not very highly correlated. But the regions discovered by our approach capture a much higher variance which is an indication that our framework successfully discovers regions with highly correlated attributes.

**Table 6.** Cumulative Variance Captured by the first k PCs in Experiment 2

| Region    | Variance Captured | Size |
|-----------|-------------------|------|
| Texas     | 57.10%            | 1655 |
| Region 48 | 84.81%            | 30   |
| Region 28 | 77.19%            | 19   |
| Region 17 | 73.15%            | 39   |
| Region 22 | 72.30%            | 16   |
| Region 32 | 70.39%            | 44   |

One could argue that any method that divides data into sub-regions increases the variance captured since lower numbers of objects are involved. This is true to some extent but additional experiments that we conducted suggest that the regions discovered by the region discovery framework are significantly better than randomly selected regions. In particular, we ran experiments where we created regions at

random and computed the variance captured for those regions. Due to space limitations, we only provide a brief summary of the results here. The highest variance captured using random regions is 72% with 16 objects whereas in our approach it is 84% with 30 objects. In general, the regional variance captured using our framework was at an average 9.2% higher than the variance captured by random regions.

#### 4.4. Post Processing via Regression Analysis Results

The post processing phase first applies regression analysis to the global data by selecting Arsenic as the dependent variable and the other 7 chemical variables as the independent variables. The OLS regression result shows that Molybdenum, Vanadium, Boron, and Silica increase the arsenic concentration, but Sulfate and Fluoride decrease it Texas-wide. Next, it retrieves the list of the top-ranked regions and applies the regression analysis to regions. The result of the global regression and one example of regional regression analysis are shown in Tables 7 and 8, respectively.

**Table 7.** Regression Result for Global Data

| As                              | Coef.   | Std.Er | t     | P> t       |
|---------------------------------|---------|--------|-------|------------|
| Mo                              | 0.101   | 0.0204 | 4.95  | 0          |
| V                               | 0.211   | 0.0048 | 43.55 | 0          |
| B                               | 0.0027  | 0.0003 | 9.49  | 0          |
| Fl                              | -0.6693 | 0.159  | -4.34 | 0          |
| SiO2                            | 0.0726  | 0.0115 | 6.3   | 0          |
| Cl                              | 0.0008  | 0.0008 | 0.97  | 0.331      |
| SiO4                            | -0.001  | 0.0007 | -1.87 | 0.062      |
| const                           | -1.696  | 0.4902 | -3.46 | 0.001      |
| R-squared –Value:               |         |        |       | 68%        |
| <b>Adjusted R-squared Value</b> |         |        |       | <b>68%</b> |

**Table 8.** Regression Result for Region 10

| As                              | Coef.  | Std.Er | t     | P> t          |
|---------------------------------|--------|--------|-------|---------------|
| Mo                              | 0.7297 | 0.2731 | 2.67  | 0.013         |
| V                               | 0.234  | 0.031  | 7.52  | 0             |
| B                               | -0.007 | 0.004  | -1.74 | 0.094         |
| Fl                              | -4.996 | 3.4254 | -1.46 | 0.156         |
| SiO2                            | -0.071 | 0.0886 | -0.8  | 0.428         |
| Cl                              | 0.0138 | 0.0071 | 2.91  | 0.066         |
| SiO4                            | -0.019 | 0.0142 | -3.34 | 0.192         |
| const                           | 7.3982 | 4.0134 | 1.84  | 0.076         |
| R-squared –Value                |        |        |       | 95.03%        |
| <b>Adjusted R-squared Value</b> |        |        |       | <b>93.73%</b> |

**Discussion:** R-Squared value is equal to 68.3% for the state of Texas, which means 68.3% of the arsenic variance can be explained by other 7 chemical variables for Texas-wide data. R-squared value increased from 68% to 93.73% in Region 10, which indicates that in this region there exist stronger correlations between arsenic and the other variables. Also globally, Chloride (Cl) and Sulfate (SiO<sub>4</sub>) are not significant as predictors for arsenic concentration; but in this region, they are significant. Conversely, Boron, Fluoride, and Silica are globally significant and highly correlated with arsenic, but this is not the case in Region 10. This information is very crucial to domain experts who seek to determine the controlling factors for arsenic pollution, as it can help to reveal hidden regional patterns and special characteristics for this region. For example, in this region, high arsenic level is highly correlated to high Sulfate and Chloride levels, which is an indication of external factors that play a role in this region, such as a nearby chemical plant or toxic waste. Our framework is able to successfully detect such hidden regional correlations.

Our approach can be viewed as using different regression function for different regions which shows similarity to the approach used in Geographically Weighted Regression (GWR) [12]. In GWR, a weight function which is a function of spatial location is used to differentiate regression functions for different locations, whereas in our work we first discover highly correlated regions maximizing a PCA-based based fitness function and then create regional regression functions for each region.

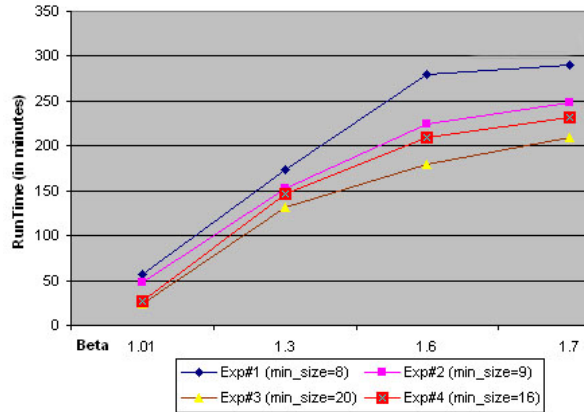
The global and regional regression results show that the relationship of the arsenic concentration with other chemical concentrations spatially varies and is not constant over space which proves the need for regional knowledge discovery. In other words, there are significant differences in arsenic concentrations in water wells across regions in Texas. Some of these differences are found to be due to the varying impact of the independent variables on the arsenic concentration. In addition, there are unexplained differences that are not accounted for by our independent variables, which might be due to external factors, such as toxic waste or the proximity of a chemical plant.

#### **4.5. Implementation Platform and Efficiency**

The components of the framework described in this paper were developed using an open-source, Java-based data mining and machine learning framework called Cougar<sup>2</sup>[6], which has been developed by our research group. All experiments were performed on a machine with 1.79 GHz of processor speed and 2GB of memory.

The parameter  $\beta$  is the most important factor with respect to run time. The run times of the experiments with respect to the  $\beta$  values used are shown in Figure 3. For example for  $\beta=1.01$ , it takes about 30 minutes to run the experiment, whereas it takes about 2 hours to run for  $\beta=1.6$ .

We observed that more than 70% of the computational resources are allocated for determining regional fitness values when discovering regions. Even though our framework repeatedly applies PCA to each explored region combination until no further improvement is made, it is still efficient compared to approaches in which PCA is applied that many times using other statistical tools.



**Fig. 3.** Run Times vs.  $\beta$  Values

## 5 Conclusion

This paper proposes a novel framework to discover regions and automatically extract their corresponding regional correlation patterns that are globally hidden. Unlike other research in data mining that uses PCA, our approach centers on discovering regional patterns and provides a comprehensive methodology to discover such patterns. The proposed framework discovers regions by employing clustering algorithms that maximize a PCA-based fitness function and our proposed post-processing techniques derive regional correlation relationships which provide crucial knowledge for domain experts. We also developed a generic pre-processing method to select the best  $k$  value for the PCA-based fitness function for a given dataset.

Additionally, a new similarity measure is introduced to estimate the structural similarity between regions based on correlation sets that are associated with particular regions. This similarity measure is generic and can be used in other contexts when two sets of objects have to be compared based on other information (e.g. eigenvectors) that has been derived from their first  $k$  principal components.

The proposed framework was tested and evaluated in a real world case study that analyzes regional correlation patterns among arsenic and other chemical concentrations in Texas water wells. We demonstrated that our framework is capable of effectively and efficiently identifying globally hidden correlations among variables along with the sub-regions that are interesting to the domain experts.

As far as the future work is concerned, we are planning to conduct extensive comparative study of our regional patterns and the co-location patterns reported in [10] for the same dataset. We are also working on developing different PCA-based fitness functions that put more emphasis on the dependent variable with the goal of developing regional regression techniques.

## References

1. Achtert, E., Böhm C., Kriegel, H.P., Kröger, P., Zimek A.: Deriving quantitative models for correlation cluster. Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, PA, , pp.4-13.c, 2006.
2. Anselin, L.: Spatial Econometrics: Methods and Models. Netherlands: Kluwer, 1988.
3. Böhm C., Keiling, K., Kröger, P., Zimek, A.: Computing Clusters of Correlation Connected Objects. In Proc. ACM SIGMOD Int. Conf. on Management of Data, Paris, France, 2004.
4. Celik, M., Kang, J., Shekhar, S., Zonal Co-location Pattern Discovery with Dynamic Parameters, In Proc. of 7th IEEE Int'l Conf. on Data Mining, Omaha, Nebraska, 2007.
5. Choo, J., Jiamthapthaksin, R., Sheng Chen, C., Celepcikay, O.U., Giusti, C., Eick, C.F.: MOSAIC: A proximity graph approach for agglomerative clustering. The 9th Int'l Conference on Data Warehousing & Knowledge Discovery (DaWaK 2007), Germany, 2007
6. Cougar<sup>2</sup> Framework, <https://cougarsquared.dev.java.net/>
7. Cressie, N.: Statistics for Spatial Data (Revised Edition). New York: Wiley, 1993.
8. Data Mining and Machine Learning Group, University of Houston, <http://www.tlc2.uh.edu/dmmlg>
9. Ding, W., Jiamthapthaksin, R., Parmar R., Jiang D., Stepinski, T., and Eick, C.F., Towards Region Discovery in Spatial Datasets, in Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), Osaka, Japan, May 2008.
10. Eick, C.F., Parmar, R., Ding, W., Stepinski, T., Nicot, J.P.: Finding Regional Co-location Patterns for Sets of Continuous Variables in Spatial Datasets, in Proc. 16th ACM SIGSPATIAL International Conference on Advances in GIS (ACM-GIS), Irvine, California, November 2008.
11. Eick, C.F., Vaezian, B., Jiang D., J.: Discovering of interesting regions in spatial data sets using supervised clustering. Proc. of the 10th European Conference on Principles of Data Mining and Knowledge Discovery, Berlin, Germany, 2006.
12. Fotheringham, A.S., Brunson, C. & Charlton, M.: Geographically Weighted Regression: The Analysis of Spatially Varying Relationships. John Wiley, 2002
13. Johnson, R.A.: Applied Multivariate Analysis. Englewood Cliffs, N.J. : Prentice Hall, c1992
14. Jolliffe I.T: Principal Component Analysis. NY Springer, 1986.
15. Openshaw, S. 1998. "Geographical data mining: key design issues". GeoComputation'99: Proceedings Fourth International Conference on GeoComputation, Mary Washington College, Fredericksburg, Virginia, USA, 25-28 July 1999.
16. Simpson E.H.: The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society, B13:238-241, 1951.
17. Texas Water Development Board, <http://www.twdb.state.tx.us/home/index.asp>
18. Turk, M., Pentland A.: Eigenfaces for Recognition, Journal of Cognitive Neuroscience, Vol. 3, No. 1, 1991.
19. Woolridge, J.: Econometric Analysis of Cross-Section and Panel Data. MIT Press, 2002, pp. 130, 279, 420-449.
20. Yu, S., Yu, K., Tresp, V., Kriegel, H.P., Wu, M.: Supervised Probabilistic Principal Component Analysis. Proceedings of the 12th ACM SIGKDD, 464-473, ACM, NY, 2006