Discovery of Geospatial Discriminating Patterns from Remote Sensing Datasets

Wei Ding* Tomasz Stepinski† Josue Salazar‡

October 12, 2008

Abstract

Large amounts of remotely sensed data calls for data mining techniques to fully utilize their rich information content. In this paper, we study new means of discovery and summarization of knowledge contained in the spatial patterns of remote sensing datasets. Several geospatial feature variables are fused together, and the vector of their values at each spatial cell is considered as a transaction to be used in association analysis. The concept of emerging patterns is applied to ascertain the variables that exert dominant influence on the distribution of a selected class variable. A new value-iteration method is introduced to optimally split the spatial domain of the selected variable into two classes. This division is used to calculate the set of patterns that are emerging with respect to the two classes; these patterns are the controlling factors—they are responsible for the spatial distribution of the class variable. A method for a concise summarization of controlling factors is introduced using a similarity measure that is custom-made for the type of patterns stemmed from remote sensing measurements. Using such a similarity measure, controlling factors are clustered providing brief description of different manners, in which the class variable is constrained by the explanatory variables. We evaluate our method in a real-world application pertaining to the density of vegetation within the continental United States. Examination of patterns related to the high vegetation cover provides a summary of data dependencies that helps to develop a better empirical model of the vegetation growth.

Keywords

Spatial Data Mining, Spatial Frequent Patterns, Emerging Patterns, Remote Sensing Datasets

1 Introduction.

Remote sensing data, pertaining to geosciences, consists of satellite observations of climate, vegetation cover, terrain topography, lithology, soil properties, etc. A large number of such datasets is available in the public domain within the framework of Geographic Information Systems (GIS). For example, PRISM [22] provides spatial datasets related to climate (precipitation, temperature, etc.) within the continental United States and incorporates measurements that has been dated from 1971. Datasets such as PRISM offer an unprecedented opportunity for studying various aspects of Earth Science and to predict and address environmental and other ecological problems. For example, understanding the spatial variation of drainage density—the density of land surface dissection by river networks—is related to the problem of assessing the risk of damage and degradation of the landscapes. Studying the spatial distribution of carbon flux, controlled by land precipitation, land and ocean temperature, and terrestrial biomass loss, leads to a better understanding of global warming. Sufficient data exist to address such problems, what is lacking is a methodology that can efficiently distill vary large amount of data into a usable knowledge. Data mining techniques are well suited to provide such methodologies. In this paper, we introduce a concept of geospatial discriminate patterns and a new similarity measure to capture and summarize complex interactions among geospatial variables.

Given a geospatial dataset classified into two binary (yes/no) classes, the goal of this paper is to discover patterns of additional (explanatory) variables that are capable of distinguishing between the two classes. Such patterns are *emerging* with respect to one of the two classes; they can be used to establish factors controlling spatial distribution of the class variable. Emerging patterns have been proposed and well studied in [6, 11, 14, 13, 15, 7, 17] as means to understand the patterns contrasting two different classes. However, not much work has been done to understand the contrasts between spatially extended classes. Generalizing the methods of standard emerging patterns to spatial domain is a non-trivial task. Geospatial data often contain continuous variables that need to be categorized in order to be subjected to association analysis. Categorization inevitably leads to information loss as it introduces sharp artificial boundaries between different regions. Furthermore, in contrast to the assumption

*University of Massachusetts Boston, ding@cs.umb.edu.
†Lunar and Planetary Institute, tom@lpi.usra.edu.
‡University of Houston-Clear Lake, SalazarJ4857@uhcl.edu

that data instances are independent in traditional data mining, spatial patterns often exhibit spatial continuity and high autocorrelation among geographically nearby features.

Our proposed methodology aims at addressing these problems. Specifically, we focus on the following three challenges: (1) identifying representative patterns of explanatory variables that capture statistical difference between geospatial classes, (2) seeking the optimal spatial boundary between classes from which the class-discriminating patterns can be derived, and (3) summarizing the identified patterns and presenting domain experts with a relevant report. To address challenges (1) and (2), we introduce the concept of *geospatial discriminating patterns* and propose a new value-iteration method designed to find the optimal geospatial boundary between classes using a reinforcement-learning model. To address challenge (3), we define a similarity measure using information theory and use the proposed similarity metric to summarize identified patterns by clustering them into a small number of "super-patterns". We design and implement a set of algorithms to efficiently mine class-discriminating patterns. We apply our methods to a real-world case study focusing on understanding the variations of vegetation density within the continental United States. Specifically, we find patterns of explanatory variables that control geographical extent of "high" density of vegetation. Examination of discovered patterns provides a summary of data dependencies that helps to develop a better empirical model of the vegetation growth.

Outline. The remainder of the paper is organized as follows. Section 2 discusses related work. Section 3 gives formal definition of the research problems. Section 4 presents the value-iteration method for optimal boundary discovery. Section 5 defines a similarity measure for patterns. Algorithm design is discussed in Section 6. We report our experimental results in Section 7 and conclude the paper in Section 8.

2 Related Work.

First introduced by Dong et al. in [6], emerging patterns are the patterns whose supports increase significantly from one dataset to another. Li et al. [11, 15] have systematically studied various statistical measures of "emergence", including relative risk ratio, odds ratio, risk difference, and delta-discriminative emerging patterns. In this paper, we adopt the relative risk ratio as the measure of pattern emergence. Emerging patterns have been applied to many scientific applications, including medical science [2, 14, 12, 15], network traffic control [5], and data credibility analysis [21], etc. For example, in medical studies, a single dataset of subjects can be divided into two mutually exclusive and exhaustive classes: subjects in a case group with a disease and those in a control group without the disease [27]. The research design of the case and control groups is to look backward in time to find what risk factors are more likely to cause the disease. Hence, the task is to mine patterns that are frequent in the case group but less frequent in the control group. The identified patterns are usually used to build a classifier to predict the disease from the presence or absence of particular symptoms. However, little work has been done with respect to analyzing emerging patterns in spatial datasets.

Identifying emerging patterns in spatial datasets has its own challenges. Geospatial variables are highly coupled through a complex chain of interactions resulting in their mutual inter-dependability. Ceci *et al.* [4] applied emergence to spatial databases where spatial interactions between different sets of spatial objects are stored in relational tables. We propose a different solution by seeking the optimal spatial boundary between the classes from which geospatial discriminating patterns are identified. Our ultimate goal is to discover a set of controlling factors that provides knowledge for building empirical models of chosen phenomena (represented by a given class variable).

Other studies indirectly related with our present work are spatial association rule mining [25, 10] and spatial co-location mining [25, 28, 9, 29]. These methods have done excellent work on discovering spatial associations or spatial features whose instances are frequently located together. Our work is to find patterns that capture statically important differences between two classes of a given geospatial variable.

3 Problem Formulation.

Geospatial variables consist of measurements acquired by means of satellite remote sensing. Different instruments on different satellites provide variables that reflect different aspects of the real world. Let $\mathcal{R} = label(xdx, ydy),\ x = 1, \ldots, N_x,\ y = 1, \ldots, N_y$, be a raster having dimensions of (N_x, N_y), covering the entire spatial extent of a dataset. The raster is an array of constituent grid cells (pixels), each having an area of $dx \times dy$. Let a geospatial dataset $\mathcal{O}$ be the fusion of all explanatory variable $\mathcal{F}_1, \ldots, \mathcal{F}_m$ and one geospatial class variable $\mathcal{CL}$, overlaying on the raster $\mathcal{R}$. $\mathcal{F}_1, \ldots, \mathcal{F}_m$ and $\mathcal{CL}$ are co-registered rasters having the same dimensions of (N_x, N_y), see Figure 1. Thus, each pixel in $\mathcal{R}$ is mapped one-to-one to an object in $\mathcal{O}$ having the form of $\{x, y; f_1, f_2, ..., f_m; c\}$, where x and y are spatial coordinates in the raster $\mathcal{R}$, feature value $f_i \in F_i$, $i = 1, \ldots, m$, and $c \in \{0, 1\}$. Each object is labeled "interesting" ($c = 1$) or not ($c = 0$) accord-

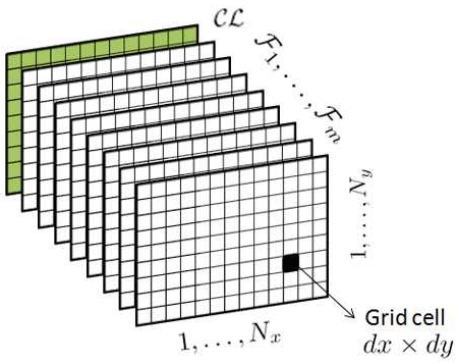

Figure 1: An illustrative example for a geospatial dataset $\mathcal{O}$.

Table 1: An example of closed frequent patterns.

TID	**Items** (3 features: $f_1,\ f_2,\ f_3$)
T_1	f_1=1, f_2=2, f_3=3
T_2	f_1=2, f_2=3, f_3=1
T_3	f_1=1, f_2=2, f_3=2
T_4	f_1=3, f_2=3, f_3=1
T_5	f_1=3, f_2=3, f_3=1

Closed Frequent Patterns	$sup(),\ \rho = 40\%$
$\{f_1 = 1,\ f_2 = 2\}$	$\frac{2}{5} = 40\%$ (T_1,T_3)
$\{f_2 = 3,\ f_3 = 1\}$	$\frac{3}{5} = 60\%$ (T_2,T_4,T_5)
$\{f_1 = 3,\ f_2 = 3,\ f_3 = 1\}$	$\frac{2}{5} = 40\%$ (T_4,T_5)

ing to its value $cl \in \mathcal{CL}$ and a user-specified threshold t_{CL}. For example, in our case study, an object is labeled $c = 1$ (high vegetation density), if its vegetation density ≥ 166.3. The dataset $\mathcal{O}$ is then classified into two mutually exclusive and exhaustive classes: dataset $\mathcal{O}_p$ including all the objects with $c = 1$ and dataset $\mathcal{O}_n$ including all the objects with $c = 0$.

Challenge 1: Given the spatial dataset $\mathcal{O}$ that contains two mutually exclusive classes $\mathcal{O}_p$ and $\mathcal{O}_n$, what are the effective representative patterns that contrast these two geospatial classes?

We seek patterns (thereafter referred to as geospatial discriminating patterns) that effectively capture statistical difference between the two classes. This approach is based on the concept of association analysis, in particular, closed patterns [20] and emerging patterns [6].

Disregarding the location information (x, y) and the class label c, each object in $\mathcal{O}$ can be viewed as a fix-length transaction containing a set of items $\{f_1, f_2, ..., f_m\}$. An itemset is a set of items. Thus, each object in the dataset $\mathcal{O}$ can be viewed as a transaction of m-itemset, which contains exactly m items. The dataset $\mathcal{O}$ can be viewed as a set of $N_x \times N_y$ transactions. A transaction is said to support an itemset if the itemset is a subset of the transaction. For example, a 5-item transaction $\{f_1 = 1, f_2 = 2, f_3 = 3, f_4 = 4, f_5 = 5\}$ supports an itemset $\{f_1 = 1, f_3 = 3\}$ or an itemset $\{f_2 = 2, f_4 = 4, f_5 = 5\}$, but not $\{f_1 = 1, f_2 = 5\}$ because the last itemset is not a subset of this 5-item transaction. An itemset is frequent if the number of transactions that support the itemset are greater than a user-specified minimum support threshold.

Definition 3.1. **(Closed Pattern).** A closed pattern is an itemset X in the dataset $\mathcal{O}$, $X = \{C \mid C \subseteq I \wedge \neg\exists C' \subseteq I,\ C \subset C',\ support(C) = support(C')\}$, where I is the set of all items in $\mathcal{O}$, and support count $support()$ refers to the number of transactions that support a particular itemset. In other words, if X is a closed pattern, none of its immediate supersets has exactly the same support count as X.

Closed patterns effectively reduce the total number of itemsets because they present minimal representation of a set of non-closed itemsets without losing their support information. The support count for the non-closed itemsets can be calculated directly from the closed itemsets.

We are interested in frequent patterns because infrequent itemsets are likely to be insignificant trivial patterns that happen by chance.

Definition 3.2. **(Closed Frequent Pattern).** A pattern X of the dataset $\mathcal{O}$ is a closed frequent pattern if X is closed and its support

$$sup(X) = \frac{support(X)}{|\mathcal{O}|} \geq \rho$$

where ρ is a user-specified support threshold.

Example. Table 1 shows a small 3-feature dataset of 5 objects. The following itemsets are closed frequent patterns, assuming the support threshold $\rho = 40\%$: $\{f_1 = 1,\ f_2 = 2\}$, $\{f_2 = 3,\ f_3 = 1\}$, $\{f_1 = 3,\ f_2 = 3,\ f_3 = 1\}$. Pattern $\{f_1 = 1\}$ is not a closed pattern because its immediate superset $\{f_1 = 1,\ f_2 = 2\}$ has exactly the same support. Even though $\{f_1 = 3,\ f_2 = 3,\ f_3 = 1\} \supset \{f_2 = 3,\ f_3 = 1\}$, both patterns are closed because their immediate supersets have different supports.

We define the *footprint* of an itemset to measure its spatial distribution.

Definition 3.3. **(Footprint).** The footprint of an itemset X, $fprint(X) = \{\Pi_R(T) \mid T \supseteq X, T \in \mathcal{O}\}$,

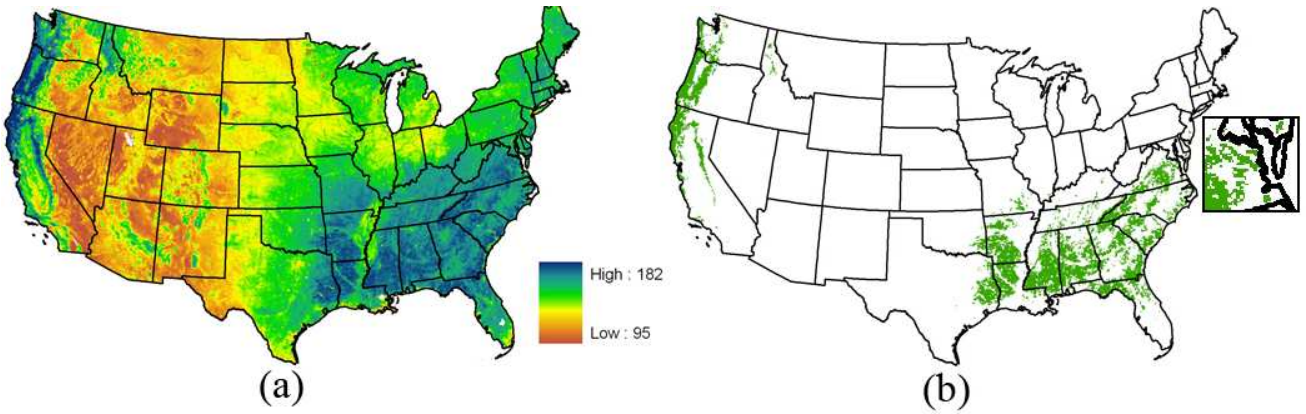

Figure 2: (a)Vegetation coverage using the Normalized Difference Vegetation Index. (b) The footprint of high vegetation region defined by a categorized class variable is shown in green. A zoomed-in window centered on the states Virginia and Maryland shows sharp unnatural boundaries of high vegetation footprint in details.

is the projection of the objects that support X into $\mathcal{R}$, the spatial reference system of the dataset $\mathcal{O}$. Function $\Pi_R(T)$ defines a pixel of R whose correspondent object, or in this case, a transaction T, supports the itemset X.

LEMMA 3.1. *The union of footprints of all closed patterns of the dataset $\mathcal{O}$ covers the entire spatial extent of $\mathcal{O}$ losslessly.*

Proof. (Proof by Contradiction.) Assume that pixel $p = \{f'_1, f'_2, \ldots, f'_m\}$ is not covered by the union of the footprints of all closed patterns of $\mathcal{O}$. If this is the case, the pattern $\{f'_1, f'_2, \ldots, f'_m\}$ is not closed, and it is not a superset of any closed patterns according to Definition 3.3. Hence, this pattern must be a proper subset of at least one closed pattern. Recall that p is not closed. In this case, p has the same support as at least one closed pattern according to Definition 3.1, otherwise p would be closed. Thus p is in the footprint of at least one closed pattern, contradicting our assumption that p is not covered by the union of the footprints of all closed patterns.

Lemma 3.1 indicates that, in addition to providing the minimal representation of a set of non-closed itemsets, closed patterns completely dominate the whole spatial extent of the dataset in a lossless way.

Our goal is to identify patterns that distinguish between geospatial classes, such patterns should possess two properties: (1) their footprints should cover a good portion of the class of interest $\mathcal{O}_p$ (good representative), and the patterns should be more frequent in $\mathcal{O}_p$ but rare in $\mathcal{O}_n$. We define *geospatial discriminating patterns* to satisfy such needs.

DEFINITION 3.4. **(Geospatial Discriminating Pattern)**. Let $\mathcal{O}$ be a dataset consisting of $\mathcal{O}_p$ with all $c = 1$ objects and $\mathcal{O}_n$ with all $c = 0$ objects, a pattern X is defined as a geospatial discriminating pattern if X is closed and its growth ratio

$$DEP^X_{\mathcal{O}} = \frac{\sup(X, \mathcal{O}_p)}{\sup(X, \mathcal{O}_n)} \geq \delta$$

where δ is a user-defined minimum growth-ratio threshold.

Combining the strength of closed patterns and emerging patterns, we submit that geospatial discriminating patterns can effective capture statistical distinctions between the two geospatial classes.

4 Discovery of Optimal Boundary.

In geospatial domain, both class and explanatory variables frequently contain real-valued entries. For example, the vegetation density shown in Figure 2(a) uses the Normalized Difference Vegetation Index, a real numeric value, to measure the amount of green vegetation in a given location. If vegetation density is the class variable, we need to classify the continental United States into two (high density, not-high density) regions using a user-specified threshold. Selecting such a threshold value is an arbitrary user choice. This creates artificial sharp boundary between those two sets which inevitably leads to information loss. Figure 2(b) depicts the spatial footprints of such datasets. Smooth transition from high to low vegetation is observed in Figure 2(a), but the footprint of the categorical high-vegetation dataset has unnatural sharp and complicated boundary in Figure 2(b).

Challenge 2: Instead of using a user-specified arbitrary threshold, how can we learn the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ that maximizes the discovery of geospatial discriminating patterns?

It is clear that good class-discriminating patterns should be significantly frequent in $\mathcal{O}_p$ and rather rare in $\mathcal{O}_n$. Hence, we can use the footprints of the best patterns to guide our learning algorithm to find the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$. We present a value-iteration method to calculate the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ using a reinforcement learning model. The method is defined by the following four components:

1. Initial State: S_0 defines an initial boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ based on an arbitrary threshold.
2. Transition Model: $Transit(s, a, s')$ denotes the probability of reaching state s' if action a is done on state s. The transitions are Markovian, that is, the probability of reaching s' from s depends only on s and not on the history of earlier states.
3. Reward Function: $Reward(s)$ defines the set of top k geospatial discriminating patterns that have the highest values of growth ratio in state s.
4. Utilities of States: $Utility(s) = Reward(s) + max_a \sum_{s'} Transit(s, a, s')Utility(s')$. $Utility(s)$ is the sum of rewards of the immediate reward for state s and the expected utility of the next state, assuming that an optimal action a is chosen. Note that $Utility(s)$ and $Reward(s)$ measure different quantities; $Reward(s)$ calculates the current reward for being in state s, whereas $Utility(s)$ is the overall total reward from s onwards.

The definition of the utilities of states, $Utility(s)$, is essentially a Bellman Equation [1] in reinforcement learning. The utilities of the states can be solved by Bellman update [1]:

$$Utility_{i+1}(s) \leftarrow Reward(s) + max_a \sum_{s'} Transit(s, a, s')Utility_i(s') \quad (4.1)$$

We start with an initial state, calculate the right-hand side of the equation, and plug it into the left-hand side—thereby updating the utility of each state from the utilities of its neighbors. The fixed point of the algorithm is the optimal boundary solution. In each state, we learn a boundary closer to the optimal solution. We repeat this until the iteration converges. It has been proved that value iteration is able to converge to a unique solution of the Bellman equation [24]. We utilize this reinforcement learning method to use observed rewards to learn the optimal boundary between the geospatial classes $\mathcal{O}_p$ and $\mathcal{O}_n$.

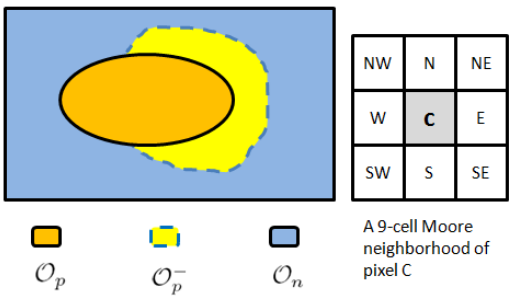

Figure 3: An example for Vote Cellular Automaton.

For $Transit(s, a, s')$, the optimal action a is executed by a complete iteration on boundary modification using the following 3 steps: we first identify top k geospatial discriminating patterns of $\mathcal{O}$; we then calculate the union of the footprints of the top k patterns; finally, we iteratively modify the boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ in current state s to produce the new boundary for $\mathcal{O}'_p$ and $\mathcal{O}'_n$ in next state s', using the footprints of the k patterns; here k is a user-specified threshold. A Cellular Automaton (CA) tool, Vote CA [8], is used to modify the boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$. Before a Vote CA can be applied, we introduce another dataset $\mathcal{O}_p^- \subset \mathcal{O}_n$ as the buffer zone for the boundary modification. Recall, t_{CL} is a user-defined threshold to determine whether an object is interesting ($c = 1$) or not ($c = 0$). $\mathcal{O}_p^-$ contains the objects whose values on cl are closest to the threshold t_{CL}. Formally, $\mathcal{O}_p^- = \{o \mid o \in \mathcal{O}_n, abs(cl(o) - t_{CL}) \leq \epsilon\}$, where the parameter ϵ control the size of the buffer zone.

We use the Vote CA as a finite state machine to "smooth-off' jagged edges. The Vote CA can be described as follows: the dataset $\mathcal{O}$ is mapped to a grid lattice having values of 0, 1, 2. Each object o in the lattice has one of the 3 values: 1 if $o \in \mathcal{O}_p$, 2 if $o \in (\mathcal{O}_p^- \cap \cup_{X_i \in top\_k\_patterns} fprint(X_i))$, and 0, otherwise. A pixel has a value 1 if it belongs to the class of interest $\mathcal{O}_p$; it has a value 0 (not an interesting value at all) if it is in $\mathcal{O}_n$ and is not included in the footprints of any top k patterns; it has a value 2 if it is in the intersection of the buffer zone $\mathcal{O}_p^-$ and the union of the footprints of the top k patterns. We will modify the boundary using the value-2 pixels. At each update of the Vote CA, the new value of a pixel is determined according to the values possessed by the nine sites in its Moore neighborhood (See Figure 3). The rules of the Vote CA are given in Table 2. A value-2 pixel will be

Table 2: A table for the voting rules of the Vote CA.

Neighborhood Pixel Value	Pixel Value	New Pixel Value
any value	1	1
any value	0	0
at least one of the 8 neighbors is 1	2	1

upgraded to a value-1 pixel (its class is changed from $\mathcal{O}_n$ to $\mathcal{O}_p$) if it has at least one value-1 pixel in its Moore neighborhood. To complete the transition from s to s', the optimal action a is that we run the Vote CA through the entire spatial extent of $\mathcal{O}$ repeatedly until no more pixels are updated. The new boundary is propagated globally by means of local updates in each pixel's Moore neighborhood.

The overall value-iteration algorithm modifies the boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ iteratively, using the footprints of the best top k geospatial discriminating patterns. The algorithm will converge when no more new top k geospatial discriminating patterns can be identified. The converged boundary is the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ with respect to the k best geospatial discriminating patterns.

5 Pattern Summarization

In the value-iteration method, a relatively large k needs to be selected to give near-complete coverage on the footprint of $\mathcal{O}_p$. In our case study experiments we use $1,500$ to $2,000$ best geospatial discriminating patterns. Once the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ are identified, it is desirable to summarize the top k patterns derived from classes $\mathcal{O}_p$ and $\mathcal{O}_n$ so the results are usable to a domain scientist. Such summarization is achieved by clustering the k patterns into a small number of "super-patterns". A distance function has to be defined beforehand in order to enable a clustering algorithm. One typical way to measure the distance is using similarity measure between the patterns and then convert it into a distance measure with $distance = \frac{1}{similarity} - 1$.

In this paper, we define a new similarity measure between patterns, based on information theory and inspired by the method proposed by Lin in [16]. Our similarity measure takes advantage of the fact that discretization of explanatory variables results in a set ordinal (rather than categorical) variables. This is because original real-valued data has a natural orientation (large-to-small).

Given two geospatial discriminating patterns X and Y of $\mathcal{O}$, we define the similarity between them as:

Table 3:
The i_{th} feature between patterns X and Y.

Cases	X	Y
Case 1	X_i	Y_i
Case 2	–	Y_i
Case 3	X_i	–
Case 4	–	–

$$s(X,Y) = \frac{\sum_{i=1}^{m} s(X_i, Y_i)}{m} \tag{5.2}$$

where X_i, Y_i is the value of i_{th} feature, f_i, of patterns X and Y, respectively. Lin in [16] defines a similarity metric in information theoretic terms, which has been proved to be effective for measuring the similarity between ordinal values. Specifically, the similarity between two ordinal values X_i and Y_i is measured by the ratio between the amount of information on the commonality of X_i and Y_i and the information needed to describe both X_i and Y_i. However, in the context of geospatial discriminating patterns, a feature f_i is not always present in both patterns. There are four possible arrangements of the presence of the i_{th} feature between patterns X and Y. Here we use "–" to denote the feature that is not present in a pattern.

For Case 1, the similarity between two ordinal values X_i and Y_i is

$$s(X_i, Y_i) = \frac{2 \times log\ P(X_i \vee Z_1 \vee Z_2 \ldots \vee Z_k \vee Y_i)}{log\ P(X_i) + log\ P(Y_i)} \tag{5.3}$$

where $P()$ is probability distribution and $Z_1, Z_2, \ldots, Z_k$ is the intervals delimited by X_i and Y_i. The commonality between two ordinal values is the interval delimited by them.

For Case 2, if feature f_i is absent in a pattern X, it is necessarily to check every value of f_i present in the footprint of X with respect to Y_i. Let feature f_i overall have n ordinal values $Z_1, Z_2, \ldots, Z_n$, we define the similarity between "–" and Y_i as

$$s(-, Y_i) = \sum_{k=1}^{n} P_X(Z_k) s(Z_k, Y_i) \tag{5.4}$$

where $P_X(Z_k)$ is the probability of value Z_k in all transactions that support pattern X. $Y_i \in \{Z_1, Z_2, \ldots, Z_n\}$ (Y_i is one of the Z's) and $\sum_{k=1}^{n} P_X(Z_k) = 1$. For a feature

dataset F_i, it is straightforward to calculate the probability of value Z_k. Notice that $P_X(Z_k) = 0$, if a Z_k does not exist in the footprint of X at all. Here $s(-, Y_i)$ is a weighted average between all ordinal values presented in the footprint of patterns X and Y_i.

Similarly, for Case 3, the similarity between X_i and "–" is

$$s(X_i, -) = \sum_{k=1}^{n} P_Y(Z_k) s(X_i, Z_k) \tag{5.5}$$

where $P_Y(Z_k)$ is the probability of value Z_k in all transactions that support pattern Y.

For Case 4, feature f_i is absent in both patterns. We check the probability distribution of all ordinal values $Z_1, Z_2, \ldots, Z_n$ in patterns X and Y, and calculate a weighted average of using a pairwise comparison

$$s(-, -) = \sum_{l=1}^{n} \sum_{k=1}^{n} P_X(Z_l) P_Y(Z_k) s(Z_l, Z_k) \tag{5.6}$$

In summary, we align geospatial discriminating patterns, calculate the similarity between every feature of $f_1, \ldots, f_m$, and we take the mean of the m similarity values as the overall similarity between the patterns.

6 Algorithm Descriptions.

We have designed and implemented the algorithms for data preprocessing, geospatial discriminating pattern mining, and pattern summarization. Figure 4 depicts the flow chart of the whole procedure. The method for data preprocessing will be discussed in Section 7. The method for similarity measure has been discussed in detail in the previous section. In this section, we present our design for the algorithm mineGDP to identify top k geospatial discriminating patterns using the proposed value-iteration method. Given two geospatial classes $\mathcal{O}_p$ and $\mathcal{O}_n$, our method seeks the optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$, using top k geospatial discriminating patterns. The algorithm consists of the following 3 steps (see Algorithm 1):

1. Mine top k geospatial discriminating patterns of $\mathcal{O}_p$ and $\mathcal{O}_n$ (lines 1-5).
2. Construct the footprints of the union of the top k patterns (lines 10-11).
3. Modify the boundary of $\mathcal{O}_p$ and $\mathcal{O}_n$ using the top k patterns until no new patterns can be generated in the k patterns (lines 3-16).

The closed-frequent-pattern generation function closedFrequentPattern-gen at line 4 uses a closed frequent itemset method introduced by Burdick et al. in [3]. A main computational part of the algorithm is at Step 2 for constructing the footprints. Instead of enumerating every single pattern of the k patterns, lines 10 and 11 in Algorithm 1 shows that we can speed up the computation by simply computing the intersection of the footprints of 1-itemsets that are represented by the top k patterns. For example, the footprint of $\{f_1 = 3, f_3 = 2\}$ is the intersection of the footprint of $\{f_1 = 3\}$ and that of $\{f_3 = 2\}$. The upper boundary of the size of 1-itemsets is determined by the total number of items in I, which is usually significantly lower than k. In addition, the top k patterns often share subpatterns with each other. The size of 1-itemsets to be computed tends to be much less than the total number of items. Another gain in performance is that we utilizes multi-thread programming. Lines 12 and 13 shows that Vote Cellular Automaton (Vote-CA) is used to iteratively modify the boundary of $\mathcal{O}_p$ and $\mathcal{O}_n$ using the footprints of the union of the top k patterns. We implement the Vote CA using multiple threads in parallel because the new value of each grid cell is determined solely by its 9 grid cells in Moore neighborhood.

7 Experimental Results.

In this section, we present the results of applying our methods to a case study featuring real geospatial data. We have constructed a fusion of several datasets that pertain to the distribution of topography, climate, and soil properties across the continental United States. Our purpose is to identify dominant factors responsible for spatial distribution of the region of high vegetation density. The datasets are summarized in Table 4. The spatial distribution of vegetation density is approximated by the distribution of the Normalized Difference Vegetation Index (NDVI). The NDVI is an index calculated from visible and near-infrared channels of satellite observations, and it serves as a standard proxy for vegetation density. The 8 explanatory variables can be divided into climate-related (average annual precipitation rate, average minimum annual temperature, average maximum annual temperature, and average dew point temperature), soil-related (available water capacity, permeability, and soil pH), and topography-related (elevation). The available water capacity is the volume of water that soil can store for plants. The pH measures the degree to which water in soil is acid or alkaline. Bulk permeability relate to the physical form of the soil. The dew temperature is an indicator of relative humidity. These datasets are from different sources and are available in different spatial resolutions. We have fused all the datasets to 9 co-registered latitude-longitude grids with a resolution of $0.5^o \times 0.5^o$. Each grid has 618×982 pixels, of which 361,882 pixels ($59.6\% = \frac{361882}{618 \times 982}$) have values for all the 9 variables.

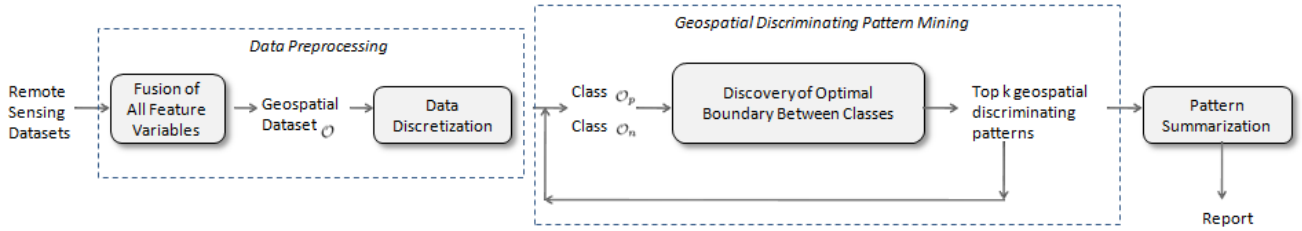

Figure 4: The flow chart of our method.

Table 4: Nine geospatial datasets used in our case study.

Variable Abbreviation	Short Description
1. tmax	Average annual maximum temperature PRISM climate mapping system [22]
2. tmin	Average annual minimum temperature PRISM climate mapping system [22]
3. dew	Average dew point temperature PRISM climate mapping system [22]
4. awc	Available water capacity ORNL for biogeochemical and ecological data [19]
5. ppt	Average annual precipitation PRISM climate mapping system [22]
6. elev	Elevation USGS National Map Seamless Server [18]
7. ph	Soil pH ORNL for biogeochemical and ecological data [19]
8. perm	Soil permeability ORNL for biogeochemical and ecological data [19]
9. aveveg	Vegetation growth average USGS National Map Seamless Server [18]

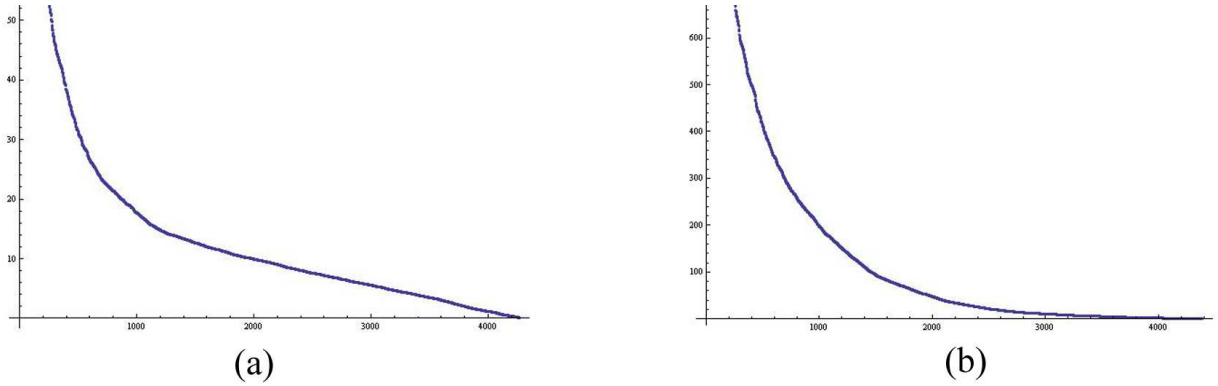

Figure 5: Growth ratio plots of geospatial discriminating patterns. (a) Geospatial discriminating patterns identified in the first iteration. (b) Geospatial discriminating patterns identified in the last iteration.

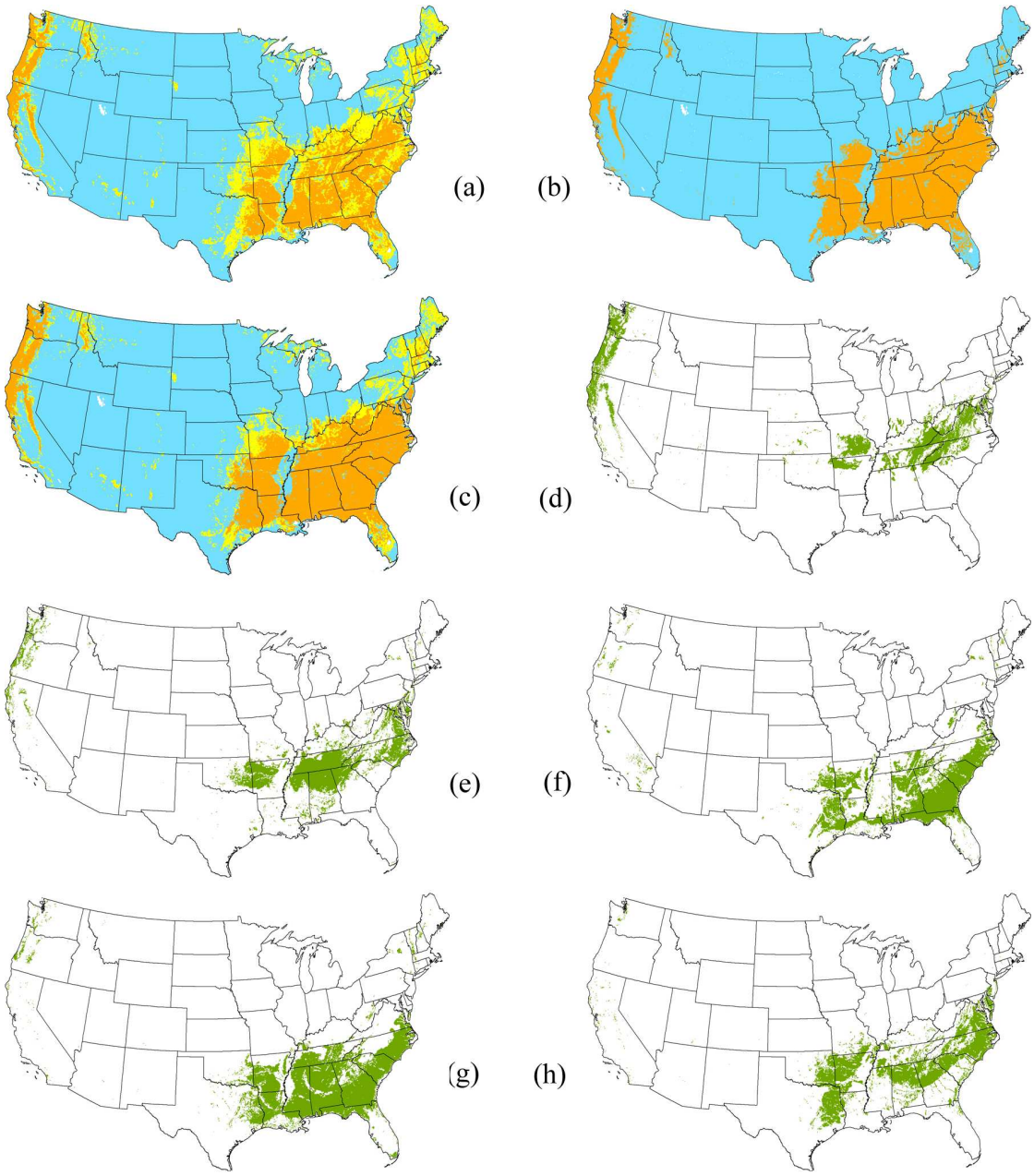

Figure 6: Experimental Results of the vegetation-cover dataset. (a) Original boundary between high vegetation cover and not-high vegetation cover. (b) Optimal boundary of high vegetation cover. (c) Optimal boundary vs. original high vegetation cover and the buffer zone. (d)-(h) Footprints of 5 groups of geospatial discriminating patterns. Color: orange - footprints of $\mathcal{O}_p$, yellow - footprints of $\mathcal{O}_P^-$, blue - footprints of $\mathcal{O}_n$, green - footprints of identified 5 super-patterns.

Algorithm 1 mineGDP: Mining geospatial discriminating patterns

Require: .
(1) Geospatial classes $\mathcal{O}_p$ and $\mathcal{O}_n$
(2) Buffer dataset $\mathcal{O}_p^-$
(3) A minimum support threshold ρ for closed frequent patterns
(4) A minimum growth-ratio threshold δ for geospatial discriminating patterns
(5) Parameter k for the top geospatial discriminating patterns

1: $old\_\mathcal{O}_p = \mathcal{O}_p$; $old\_\mathcal{O}_n = \mathcal{O}_n$;
2: pre_kCandidateset = $\emptyset$;
3: **loop**
4: candidateSet = closedFrequentPattern-gen($\mathcal{O}_p$, ρ); {Mine closed frequent patterns in $\mathcal{O}_p$ using the minimum support threshold ρ}
5: kCandidateSet = pattern-gen(candidateSet, $\mathcal{O}_p$, $\mathcal{O}_n$, δ) {Identify top k geospatial discriminating patterns using the minimum growth-ratio threshold δ}
6: **if** pre_kCandidateset == kCandidateSet **then**
7: **return** $\mathcal{O}_p$ and $\mathcal{O}_n$, and the k geospatial discriminating patterns.
8: **else**
9: pre_kCandidateset = kCandidateSet; {Remember the k geospatial discriminating patterns identified in the current iteration}
10: $CF_1 = \cup$ oneItemSet-gen(candidateSet); {Identify all unique 1-itemsets represented by the closed frequent patterns}
11: fprint_kCandidateset = $\cap$ fprint(CF_1); {Construct the footprints of the union of the top k geospatial discriminating patterns}
12: M = fprint-gen($\mathcal{O}_p$, $\mathcal{O}_n$, $\mathcal{O}_p^-$, fprint_kCandidateset); {Construct a raster M. Each grid cell has one of the 3 values: 1 if $o \in \mathcal{O}_p$, 2 if $o \in \mathcal{O}_p^- \cap$ fprint_kCandidateset, and 0, otherwise}
13: M = Vote-CA (M);
14: $\{\mathcal{O}_p, \mathcal{O}_n\}$ = boundary-change($old\_\mathcal{O}_p$, $old\_\mathcal{O}_n$, M); {Modify the boundary of $old\_\mathcal{O}_p$ and $old\_\mathcal{O}_n$ according to M}
15: **end if**
16: **end loop**

Table 5: Statistics of the nine geospatial datasets.

Dataset	Mean	Median	STD	S_n
1. tmax	1823.6	1778.0	543.4	584.3
2. tmin	461.5	414.0	535.8	579.6
3. dew	392.4	349.0	550.5	590.3
4. awc	13.3	12.0	8.1	4.8
5. ppt	80490.7	76359.0	45656.0	48913.3
6. elev	778.4	472.0	729.8	524.7
7. ph	6.49	6.57	1.15	1.16
8. perm	7.66	5.13	6.98	3.35
9. aveveg	143.8	143.0	14.6	15.5

Data Preprocessing. All the co-located datasets are subjected to a categorization procedure. Z-score transformation is a widely used method in geospatial domain. For example, Tan and Kumar et al. in [26] calculate the z-score of Earth Science time series data by subtracting off the monthly mean and dividing by the monthly standard deviation. However, a closer examination of the 9 datasets used in our case study indicates that not all the 9 datasets have the bell-shaped distributions suitable to the z-score transformation. Table 5 shows the statistics of the 9 datasets. The numerical values of the explanatory variables come from their respective distributions having quite different functional forms. The difference between columns Median and Mean indicates how off the center distribution is from the location of the bulk of the data. The difference between columns STD (standard deviation) and S_n indicates whether variables has skewed distribution of their values. S_n is introduced in [23] as a typical distance for symmetric and asymmetric distributions to measure how far away observations are from a central value. The

S_n estimator is :

$$S_n = 1.1926\ med_i\{med_j|x_i - x_j|\} \tag{7.7}$$

where *med* is the median operator. Given a set of numbers $\{x_1, \ldots, x_n\}$, for each x_i, we compute the median of $\{|x_i - x_j|, j = 1, \ldots, n\}$ to yield n numbers, then the median of the n numbers gives estimator S_n. The dataset statistics show that all datasets have more or less skewed distributions. We decide to use the standard K-means clustering algorithm as this algorithm is typically applied to real-valued objects. K-means identifies natural break points by picking the class breaks that best group similar values and maximize the differences between classes. In our experiments, we classify each dataset to 7 classes labeling the cluster from the minimum-centroid as 1 to the maximum-centroid as 7. Except for the class variable aveveg, the 8 datasets are transformed into categorical datasets containing values from 1 to 7. The vegetation density dataset aveveg is initially divided into two subsets, $\mathcal{O}_p$ with $c = 1$ (combined categories 6 and 7) and $\mathcal{O}_n$ with $c = 0$ (combined categories 1 to 5) before the mineGDP algorithm is applied. Figure 6(a) shows the footprints of $\mathcal{O}_p$ in orange, $\mathcal{O}_n$ in blue. $\mathcal{O}_p^- \subset \mathcal{O}_n$ is in yellow, which is in category 5, the closest category to categories 6 and 7 of vegetation density.

Boundary Optimization. The optimal boundary between $\mathcal{O}_p$ and $\mathcal{O}_n$ are depicted in Figure 6(b). Figure 6(c) overlays the footprints of the new $\mathcal{O}_p$ with the original $\mathcal{O}_p$ and $\mathcal{O}_p^-$. As illustrated in the figure, the boundary is expanded in the buffer zone of $\mathcal{O}_p^-$, but it does not exactly overlay the buffer zone. In our experiments, the value-iteration algorithm converges in the 4th iteration when no new patterns can be identified. Using a support threshold of 1%, top 1,500 out of 4,267 geospatial discriminating patterns are selected in the first iteration. Figure 5(a) plots the growth ratio of all 4,267 patterns. Notice that 20 patterns have growth ration of infinity. This means that those patterns are only supported in the dataset $\mathcal{O}_p$. $k = 1500$ is determined visually as the elbow position of the growth-ratio plot. Starting from second iteration, we choose a higher value $k = 2000$ because the elbow position shifts to the right after new patterns are added into the group. The value-iteration algorithm converges quickly, identifying 1176, 616, and 7 new patterns in each iteration, respectively. Figure 5(b) plots the growth ratio of all 4,381 patterns in the last iteration. The top geospatial discriminating patterns, derived from the optimized split between $\mathcal{O}_p$ and $\mathcal{O}_n$, have significantly higher growth ratio than the top patterns derived from an initial, arbitrary boundary. This is exactly what we have expected because the boundary is optimized using those top patterns.

Pattern Summarization. We classify the top 2,000 emerging patterns into 5 groups of super-patterns using K-means clustering algorithm. Figures 6(d-h) depict the footprints of the 5 super-patterns. The super-patterns represent five different major combinations of controlling factors that lead to high vegetation density; high vegetation density is associated with different factors in different spatial locations. Each super-pattern can be succinctly described on the basis of its constituent patterns. For example, the super-pattern depicted on Figure 6(g) represents high values of temperature and humidity and low values of elevation, whereas the super-pattern depicted on Figure 6(d) represents only average values of temperature and humidity but higher values of elevation. Both combinations are apparently compatible with high vegetation density, but they occur in different geographical locations. The results conform to the domain knowledge of the climate and soil conditions that support high density of vegetation. Overall, our case study shows that the range of patterns supporting high vegetation density is not completely separated in the spatial domain as is made clear from overlaps of footprints shown on Figures 6(d-h). The results indicate that it does not exist highly nonlinear dependence of vegetation density on its controlling factors. Examination of patterns related to the high vegetation cover provides a summary of data dependencies that helps to develop a better empirical model of the vegetation growth.

8 Conclusion.

In this paper, we have formulated the problem of mining geospatial discriminating patterns in the domain of geoscience. This domain uses remote sensing datasets that are mostly in the form of spatially co-registered rasters, which exhibits complex interactions among multiple attributes. We propose a value-iteration method gearing to identify the optimal boundary between geospatial classes, thus maximizing the patterns to be identified. We introduce a new similarity metric that is specially designed for ordinal variables. Discovered patterns conform to existing knowledge about the types of climates and soils that are inductive to high density of vegetation, and they deliver this knowledge in a quantitative, as well as comprehensive and systematic manner.

9 Acknowledgments.

The work is supported in part by NSF Grant IIS-0430208 and the Institute for Space Systems Operations at Houston, Texas. A portion of this research was conducted at the Lunar and Planetary Institute, which is operated by the USRA under contract CAN-NCC5-

679 with NASA.

References

[1] R. Bellman. *Dynamic Programming*. Princeton University Press, 1957.

[2] A. L. Boulesteix, G. Tutz, and K. Strimmer. A cart-based approach to discover emerging patterns in microarray data. *Bioinformatics*, 19(18):2465–72, 2003.

[3] D. Burdick, M. Calimlim, and J. Gehrke. Mafia: A maximal frequent itemset algorithm for transactional databases. In *Proceedings of the 17th International Conference on Data Engineering*, Heidelberg, Germany, April 2001.

[4] M. Ceci, A. Appice, and D. Malerba. Discovering emerging patterns in spatial databases: A multi-relational approach. In *Knowledge Discovery in Databases: PKDD 2007, Series: Lecture Notes in Artificial Intelligence*, volume 4702, pages 390–397, Berlin, Germany, 2007. Springer.

[5] G. Cormode and S. Muthukrishnan. What's new: Finding significant differences in network data streams. In *IEEE INFOCOM*, 2004.

[6] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 43–52, San Diego, California, United States, 1999.

[7] H. Fan and K. Ramamohanarao. Fast discovery and the generalization of strong jumping emerging patterns for building compact and accurate classifiers. *IEEE Transactions on Knowledge and Data Engineering*, 18:1041–4347, 2006.

[8] H. Gutowitz. *Cellular Automata: Theory and Experiment*. Bradford Books, 1991.

[9] Y. Huang, J. Pei, and H. Xiong. Mining co-location patterns with rare events from spatial data sets. *Geoinformatica*, 10(3):239–260, 2006.

[10] K. Koperski and J. Han. Discovery of spatial association rules in geographic information databases. In M. J. Egenhofer and J. R. Herring, editors, *Proceedings of the 4th Intl. Symp. Advances in Spatial Databases*, volume 951, pages 47–66, 6–9 1995.

[11] J. Li, G. Liu, and L. Wong. Mining statistically important equivalence classes and delta-discriminative emerging patterns. In *Proceedings of 13th International Conference on Knowledge Discovery and Data Mining*, pages 430–439, San Jose, California, 2007. 12-15 August.

[12] J. Li, H. Liu, S.-K. Ng, and L. Wong. Discovery of significant rules for classifying cancer diagnosis data. *Bioinformatics*, 19:ii93–ii102, 2003.

[13] J. Li and K. Ramamohanarao. The space of jumping emerging patterns and its incremental maintenance algorithms. In *Proceedings of 17th International Conf. on Machine Learning*, pages 551–558. Morgan Kaufmann, San Francisco, CA, 2000.

[14] J. Li and L. Wong. Structural geography of the space of emerging patterns. *Intelligent Data Analysis*, 9(6):567–588, 2005.

[15] J. Li and Q. Yang. Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. *IEEE Transactions on Information Technology in Biomedicine*, 11:544–552, 2007.

[16] D. Lin. An information-theoretic definition of similarity. In *International Conference on Machine Learning*, Madison, Wisconsin, July 1998.

[17] E. Loekito and J. Bailey. Fast mining of high dimensional expressive contrast patterns using zero-suppressed binary decision diagrams. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006.

[18] National Map Seamless Server. *http://seamless.usgs.gov/*, 2008.

[19] Oak Ridge National Laboratory Distributed Active Archive Center Data Holdings. *http://daac.ornl.gov/holdings.html*, 2008.

[20] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *ICDT '99: Proceedings of the 7th International Conference on Database Theory*, pages 398–416, 1999.

[21] R. Podraza and K. Tomaszewski. KTDA: Emerging patterns based data analysis system. In *XXI Fall Meeting of Polish Information Processing Society*, pages 213–221, 2005.

[22] PRISM (Parameter-elevation Regressions on Independent Slopes Model) Climate Mapping System Products Matrix. *http://www.prism.oregonstate.edu/products/matrix.phtml*, 2008.

[23] J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. American Stat. Association*, 88:1273–1283, 1993.

[24] S. Russell and P. Norvig. *Artificial Intelligence A Modern Approach*. Prentice Hall, 2003.

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

[26] M. Steinbach, P.-N. Tan, V. Kumar, S. Klooster, and C. Potter. Discovery of climate indices using clustering. In *KDD '03: Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2003.

[27] Wassertheil-Smoller. *Biostatistics and Epidemiology A Primer for Health and Biomedical Professionals*. Springer Verlag, 2004.

[28] H. Xiong, S. Shekhar, Y. Huang, V. Kumar, X. Ma, and J. S. Yoo. A framework for discovering co-location patterns in data sets with extended spatial objects. In *SDM'04: SIAM International Conference on Data Mining*, 2004.

[29] J. S. Yoo and S. Shekhar. A join-less approach for mining spatial co-location patterns. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 18, 2006.