Geoinformatica manuscript No.
(will be inserted by the editor)

Controlling Patterns of Geospatial Phenomena

Tomasz F. Stepinski · Wei Ding · Christoph F. Eick ·

Received: date / Accepted: date

Abstract Modeling spatially distributed phenomena in terms of its controlling factors is a recurring problem in geoscience. Most efforts concentrate on predicting the value of response variable in terms of controlling variables either through a physical model or a regression model. However, many geospatial systems comprises complex, nonlinear, and spatially non-uniform relationships, making it difficult to even formulate a viable model. This paper focuses on spatial partitioning of controlling variables that are attributed to a particular range of a response variable. Thus, the presented method surveys spatially distributed relationships between predictors and response. The method is based on association analysis technique of identifying emerging patterns, which is extended in order to be applied more effectively to geospatial data sets. The outcome of the method is a list of spatial footprints, each characterized by a unique "controlling pattern"—a list of specific values of predictors that locally correlate with a specified value of response variable. Mapping the controlling footprints reveals geographic regionalization of relationship between predictors and response. The data mining underpinnings of the method are given and its application to a real world problem is demonstrated using an expository ex-

Tomasz F. Stepinski
Lunar and Planetary Institute
Houston, TX 77058
E-mail: tom@lpi.usra.edu

Wei Ding
Department of Computer Science
University of Massachusetts Boston
Boston, MA 02125-3393 E-mail: ding@cs.umb.edu

Christoph F. Eick
Department of Computer Science
University of Houston
Houston, TX 77004
E-mail: ceick@uh.edu

ample focusing on determining variety of environmental associations of high vegetation density across the continental United States.

Keywords Predictors-Response Relationship, Association Analysis, Mapping Predicting Relationship, Vegetation Density, Data Mining

1 Introduction

A common problem in geoscience is to model an observed phenomenon in terms of its likely predictors. Because physical models are usually difficult to formulate, the data-centric regression modeling approach is prevalent. The most popular regression models are global—a single predictive formula is holding over the entire data space. However, in the geospatial context, where variables are often mutually interdependent in a spatially distributed fashion, assembling a viable global model can be very difficult. One solution is to construct a model in the form of a regression tree [30], which models different subsets of the data set by different linear regressions. This approach is frequently used in geoscience. For example, in studying contamination of soil by heavy metals, environmental engineers attempt to model occurrence of contaminant on factors such as soil parameters, and catchment properties [18,15,28]. Similarly, hydrologists are modeling [26,24] concentration of nitrate and/or phosphorus in stream water in terms of factors such as antecedent precipitation index, air and water temperatures, amount of discharge etc. Biologist model [11,10] occurrence of different species of plants based on co-occurrence of environmental factors such as climate, landform, soil type, geomorphology and level of urbanization, and, in agriculture, the crop yields are modeled [16] in terms of climate and soil properties. In general, these (and similar) studies have two different but related goals: (1) to obtain a classifier to be applied for objects for which the values of predictors are know but the values of response are not, (2) to accept or reject a regression model as a good description of the relationship between predictors and response. However, it has been pointed out [29] that the regression tree model can also be used for mapping spatial distribution of objects on the basis of the leaves of the tree to which they belong. The resulting map reveals a geographic partition of the data set on the basis of differences in relationships between predictors and response. For example, in [29] it was demonstrated that the same level of richness of native species of birds in Oregon is associated with different set of predictors (different leaves in the regression tree) mapped to different geographical parts of Oregon. Thus, such methodology can reveal a diversity of predictors leading to the same response.

In this paper we propose a new, fundamentally different approach to the discovery of geographic regionalization of relationships between predictors and response. Presented method follows our earlier efforts [7,27] to develop means for studying relationships between various spatial variables using a branch of data mining techniques called association analysis. The goal of association analysis is to discover rules (or patterns) that specify affinity of objects in a data set. Association analysis was originally developed [1] to analyze so-called

market basked transactions but was since applied to other domains, such as bioinformatics, medical diagnosis, and Web mining. Because of its origins, the association analysis works natively on categorical, non-spatial variables and needs to be modified for applications in geoscience. On the other hand, the analysis is completely data-centric, it requires no modeling assumptions, and is capable of yielding results based on simple, clearly understood principles.

Our core idea is as follows. First, we assemble a geospatial data set where objects are the pixels carrying local values of predictor and response variables. Throughout this paper we will also refer to predictors as "explanatory variables," and to response as "class variable." Second, we mine this data set for frequent *emerging* patterns. Note that in the context of this paper a *pattern* refers to a composition of non-spatial attributes (explanatory variables) and not to a spatial object which we refer to as a *footprint*. Initially introduced by [9], emerging patterns are the patterns that contrast two different data classes; they are frequent in one data class but rare in another data class. In our context, we divide the range of response variable into "interesting" (referred to as *phenomenon*) and "other" and mine for emerging patterns with respect to the phenomenon. Such emerging patterns are *controlling* patterns of the phenomenon because they are associated predominantly with its presence. Different patterns represent different combinations of predictors leading to a phenomenon. Mapping the footprints of controlling patterns revels regionalization and diversity of predictors leading to the same phenomenon.

The rest of the paper is organized as follows. Section 2 gives a brief introduction into emerging patterns. In Section 3 we give a detailed description of our methodology. This includes spatially-specific extensions to a standard technique of association analysis and methods for pre-processing original data to categorical form required by the association analysis. In Section 3 we present an application of our methodology to an expository example pertaining to environmental correlates of high vegetation density across the continental United States. Conclusions and future research directions are given in Section 5.

2 Emerging Patterns

Emerging patterns are patterns whose supports increase significantly from one data set to another [9], hence they are a useful tool to capture a contrast between two different data sets or between two classes within a single data set. Emerging patterns have been successfully applied in many domains including medical science [2,13,12,14], monitoring network traffic [5], data credibility analysis [20], etc. For example, in medical studies, a single data set of subjects ($\mathcal{R}$) can be divided into two mutually exclusive and exhaustive classes: subjects showing symptoms of a disease ($\mathcal{D}$) and those in a control group ($\mathcal{C}$) free from the symptoms. The patterns to be considered consist of various factors that can potentially lead to the development of the disease. The task is to mine for emerging patterns that are frequent in $\mathcal{D}$ but absent or significantly less frequent in $\mathcal{C}$. Taken collectively, the emerging patterns identify all com-

binations of risk factors leading to disease development and thus contribute to the understanding of the root causes of the disease.

Our method is based on an analogous application of emerging patterns but to the spatial data sets. The previous application of emerging pattern to spatial data sets [4] concentrated on a small number of objects making possible to store their spatial interactions in a relational table. In this paper we apply emerging patterns to raster data sets which call for a different approach. First, because most geospatial rasters contain continuous variables, these variables need to be categorized in order to be subjected to association analysis. We propose a categorization procedure that can accommodate non-Gaussian distributions and assures that all variables are categorized in a correspondent manner. Second, we propose a modification to the standard definition of pattern support in order to alleviate arbitrariness in dividing the class variable on the basis of discretization.

3 Methodology

Modern geospatial data sets originate from remote sensing and are given in the form of rasters. In such context, an *object* in a database is a raster cell or a pixel. Hereafter, we will use the terms objects and pixels interchangeably. Older data sets originate from manual measurements and are given in the form of tables. In such data sets, an object is a table entry. Our method works equally well with raster or table data sets; if necessary we perform a table to raster conversion. In most cases the measurements are real numbers from a continuous domain. Because the emerging-pattern technique requires categorical data, the first step in our method is to categorize the data.

3.1 Data Preprocessing

The numerical values of explanatory variables come from their respective distributions that could have quite different functional forms. Therefore, it is necessary to normalize the values of different variables to the common meaning. The two most important properties of any distribution is its center (μ) that indicates location of the bulk of the data, and the scale (σ) that indicates dispersion around the center. For variables having bell-shaped distributions, μ and σ can be easily estimated using the mean and the standard deviation, respectively. However, the mean and the standard deviation are biased estimates of μ and σ for variables with skewed distribution of their values. For μ, a robust estimator is the trimmed mean calculated by discarding a certain percentage of the lowest and the highest values. Note that the median, $\tilde{x}$, is a particular example of the trimmed mean. For σ, a robust estimator is a function S_n introduced by [23]:

$$S_n = c\ med_i\{med_j|x_i - x_j|\} \tag{1}$$

where c is a constant having the value of 1.1926, and med is the median operator. S_n, whose numerical complexity is $O(n \log n)$ [23], is a measure of dispersion around the center that works equally well for symmetric as well as asymmetric distributions.

In order to normalize the values of explanatory variables to a common meaning, we transform them to their modified z-scores $(x - \tilde{x})/S_n$. We refer to this expression as the "modified" z-score because it has the same form as an ordinary z-score, but with the values of the mean and the standard deviation replaced by the values of $\tilde{x}$ and S_n. Thus, the modified z-score is the number of S_n that a given value of a variable is above or below the median calculated from the global distribution of this variable. The positive values of z-score indicate upward deviations from the median, whereas the negative values of z-score indicate downward deviations from the median. Two different variables with the same z-score are "equal" in the sense that both are deviated by the same relative amount from the centers of their distributions. The data sets are categorized by first transforming them to their modified z-scores and then assigning the z-scores into n bins using $n-1$ split points. The z-score of a given variable is converted to an integer number corresponding to the bin to which it belongs. This transforms all real-valued data sets into categorical data sets with a common range.

3.2 Problem Definition

The fusion of all data sets relevant to a given task results in a geospatial data set $\mathcal{R}$—a raster where each pixel is an object having a form of a tuple

$$r = \{x, y; a_1, a_2, ..., a_m; cl\} \tag{2}$$

where the first two entries (x, y) are spatial coordinates, the next m entries $a_1, a_2, ..., a_m$ are categorical values of m explanatory variables that can potentially exert control over the class variable, and the last entry cl is a binary variable that indicates whether the class variable has a value of interest ($cl = 1$) or not ($cl = 0$). Disregarding the location information (x, y), each object in $\mathcal{R}$ can be viewed (from the point of view of association analysis) as a transaction $\{a_1, a_2, ..., a_m; cl\}$. All transactions are classified into two mutually exclusive and exhaustive sets: data set $\mathcal{D}$ grouping transactions with $cl = 1$ (phenomenon) and data set $\mathcal{C}$ grouping transactions with $cl = 0$. A pattern (itemset) is a set of items contained in a transaction. For example, assuming $m = 10$, $P = \{2, \_, \_, \_, 3, \_, \_, \_, \_, \_; 1\}$ is a pattern indicating that $a_1 = 2$, $a_5 = 3$, and $cl = 1$ while the values of all other variables are not specified. A transaction supports the pattern P if it has specified values of indicated attributes. The footprint of the pattern P is the set of pixels corresponding to transactions that support the pattern. For example, Fig. 1b illustrates the footprint of a pattern $\{\_, \_, \_, \_, \_, \_, \_, \_, \_, \_; 1\}$ where $cl = 1$ indicates the class of "high" values of vegetation density.

We defined a controlling pattern as:

Definition 1 A **controlling pattern (CP)** P in $\mathcal{D}$ is an itemset such that its growth ratio $CP_P^{\mathcal{D}}$ fulfills the criterion

$$CP_P^{\mathcal{D}} = \frac{\sup(P, \mathcal{D})}{\sup(P, \mathcal{C})} \geq \rho$$

where ρ is a user-defined minimum growth-ratio threshold, and $\sup(P, \mathcal{D})$ and $\sup(P, \mathcal{C})$ are the support of a pattern P in $\mathcal{D}$ and $\mathcal{C}$, respectively. Pattern support is basically a measure of the size of its footprint; we give a formal definition of sup() in the following subsection. We refer to such itemsets as controlling patterns, because they correspond to particular values of certain explanatory variables that happen to be associated in disproportionally large numbers with $cl = 1$ objects. It is therefore expected that they constitute controlling factors for the distribution of $cl = 1$ objects.

3.3 Calculating Pattern Support

Let's $\mathcal{F}$ be a set of transactions (corresponding to a set of pixels) which support a given pattern P, and $\mathcal{G}$ be a set of transactions (corresponding to the remaining set of pixels) which do not support P. We define the following sets:

- $\mathcal{D}_+ = \mathcal{D} \cap \mathcal{F}$, pixels of interest that support P
- $\mathcal{D}_- = \mathcal{D} \cap \mathcal{G}$, pixels of interest that do not support P
- $\mathcal{C}_+ = \mathcal{C} \cap \mathcal{F}$ pixels of no interest that support P
- $\mathcal{C}_- = \mathcal{C} \cap \mathcal{G}$ pixels of no interest that do not support P

The support of P in data sets $\mathcal{D}$ and $\mathcal{C}$ is defined as:

$$\sup(P, \mathcal{D}) = \frac{|\mathcal{D}_+|}{|\mathcal{D}|}, \ \sup(P, \mathcal{C}) = \frac{|\mathcal{C}_+|}{|\mathcal{C}|} \tag{3}$$

Thus,

$$CP_P^{\mathcal{D}} = \frac{\sup(P, \mathcal{D})}{\sup(P, \mathcal{C})} = \frac{|\mathcal{D}_+|/|\mathcal{D}|}{|\mathcal{C}_+|/|\mathcal{C}|} \tag{4}$$

where $|\ |$ denotes the number of elements in a set. Notice that $|\mathcal{D}_+| + |\mathcal{D}_-| + |\mathcal{C}_+| + |\mathcal{C}_-| = |\mathcal{R}|$. Discovering controlling patterns is a matter of evaluating $CP_P^{\mathcal{D}}$ given by Eqn. 4 for a set of viable patterns and selecting those patterns that have $CP_P^{\mathcal{D}} \geq \rho$.

Without loss of generality our method models one aspect of the class variable at the time. Thus, the set $\mathcal{D}$ is defined by an arbitrary threshold (or thresholds) that identifies this aspect. For example, in the case study of the vegetation density, $\mathcal{D}$ may correspond to pixels having "high" density of vegetation. In our application "high" density of vegetation is defined by belonging to the two highest z-score categories (see Section 4). Fig. 1(a) depicts the distribution of vegetation density in the United States. Fig. 1(b) depicts a footprint of $\mathcal{D}$ corresponding to a high density of vegetation. Note that a smooth transition from high to low vegetation density is observed in Fig. 1(a), but the

footprint of $\mathcal{D}$ in Fig. 1(b) has artificially sharp and complicated boundary—an artifact of data discretization.

In order to offset the effects of data discretization, we redesign a definition of pattern support without changing the footprints of the patterns themselves. We observe that due to the spatial continuity of class variable data, many pixels nearby the footprint of $\mathcal{D}$ are expected to have values of class variable, that although lower than a required threshold, are nevertheless quite close to this threshold. We propose to incorporate this observation into a new definition of $\sup(P, \mathcal{D})$. Specifically, we propose a following modification of $|\mathcal{D}_+|$ that we denote by $|\mathcal{D}_+|^*$:

$$|\mathcal{D}_+|^* = \sum_{o \in \mathcal{F}} w(o, \mathcal{D}) \tag{5}$$

where o is a pixel belonging to $\mathcal{F}$ and $w(o, \mathcal{D})$ is a weight determined on the basis of spatial proximity of this pixel to $\mathcal{D}$. The weight is calculated using the following formula:

$$w(o, \mathcal{D}) = \begin{cases} 1 & o \in \mathcal{D} \\ \Psi(h(o, \mathcal{D})) & o \in \mathcal{C} \end{cases} \tag{6}$$

The influence function Ψ determines the weights for objects outside the footprint of $\mathcal{D}$. In the traditional definition of pattern support, $\Psi() = 0$ and $|\mathcal{D}_+|^* = |\mathcal{D}_+|$. However, when working with spatially extended data, an influence function $\Psi() \neq 0$ better captures the character of the data; in this paper, we use a half normal distribution as the influence function:

$$\Psi(\xi) = exp(\frac{-\theta^2 \xi^2}{\pi}) \tag{7}$$

where θ, $\theta \in [0, \infty)$, is a free parameter. The function Ψ determines the weights for the objects that are within the footprint of the pattern but outside the footprint of $\mathcal{D}$; the farther a pixel o is from $\mathcal{D}$, the less it counts toward the pattern support. The function $h(o, \mathcal{D})$ used by the influence function Ψ is a special case of Hausdorff distance [17] in the spatial domain. It measures the distance between a pixel o and the data set $\mathcal{D}$ as the distance between o and the the nearest pixel in $\mathcal{D}$.

Fig. 2 illustrates the concept of weights as applied not to a particular pattern P but rather to the entire data set ($\mathcal{F}$ in Eqn. 5 is replaced by $\mathcal{R}$). The vegetation density data is used. The high vegetation density regions ($\mathcal{D}$) are surrounded by nearby pixels that have weights decreasing with increasing distance from the region. The smaller the value of θ, the more surrounding pixels will be taken into account. Thus, using our measure of support, the support of the pattern P "in" $\mathcal{D}$ is increased ($|\mathcal{D}_+|* > |\mathcal{D}_+|$) if a significant number of pixels close to the footprint of $\mathcal{D}$ conform to this pattern. Simultaneously, in such a situation, the support of P in $\mathcal{C}$ must be decreased by the same amount. This is captured by a modification of $|\mathcal{C}_+|$ by a measure denoted by $|\mathcal{C}_+|^*$

$$|\mathcal{C}_+|^* = \sum_{o \in \mathcal{F}} [1 - w(o, \mathcal{D})] \tag{8}$$

Finally, the controlling patterns are mined using the new definition of pattern support,

$$CP_P^{\mathcal{D}^*} = \frac{|\mathcal{D}_+|^*/|\mathcal{D}|}{|\mathcal{C}_+|^*/|\mathcal{C}|} \tag{9}$$

Note that in Eqn. 9, we do not modify the values of $|\mathcal{D}|$ and $|\mathcal{C}|$, and the total number of pixels is preserved: $|\mathcal{D}_+|^* + |\mathcal{D}_-| + |\mathcal{C}_+|^* + |\mathcal{C}_-| = |\mathcal{R}|$.

3.4 Measuring Spatial Aggregation of Footprints

Spatial character of controlling-pattern footprints differs from one pattern to another; patterns with more aggregated footprints are arguably more likely to reveal important controlling factors than the patterns with more disperse footprints because a pattern extending over a well-defined region is more likely to correspond to coordination of factors due to specific regional conditions whereas a pattern extending over dispersed pixels is more likely to correspond to coordination due to coincidence. We use Ripley's K function [6], a statistical method frequently applied to point pattern analysis, to quantify degree of footprint aggregation. Without considering edge effects, Ripley's K function is estimated as [6]:

$$\hat{K}(d) = \frac{F}{N^2} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} I_d(d_{ij}) \tag{10}$$

where N is the number of pixels in the footprint of the pattern P, d_{ij} is the distance between the i^{th} and j^{th} pixel, $I_d(d_{ij})$ is the indicator function which is 1 if $d_{ij} \leq d$ and 0 otherwise, F is the area of a footprint, and d is a free parameter corresponding to a distance scale. In order to infer clustering properties of a footprint, the value of $\hat{K}(d)$ is compared to the value calculated for a completely random (homogeneous Poisson process) ensemble of points that is $K_{\mathrm{o}}(d) = \pi d^2$,

$$k_P = \sqrt{\frac{\hat{K}(d)}{K_{\mathrm{o}}(d)}} = \frac{\sqrt{\frac{\hat{K}(d)}{\pi}}}{d} \tag{11}$$

where $k_P > 1$ indicates spatial aggregation, and $k_P < 1$ indicates spatial segregation. The larger value of k_P indicates a more aggregated pattern P.

3.5 Algorithm for Discovering Controlling Patterns

We have designed and implemented an algorithm called mineCP to identify the controlling patterns. We mine for controlling patterns from amongst all the patterns that are *frequent* in $\mathcal{D}$. In the first step, the mineCP algorithm (see Algorithm 1) finds these frequent patterns using a user-defined support threshold. We use an efficient depth-first search method introduced by Burdick et al. in [3] to find the frequent patterns. In the second step, the algorithm calculates the modified pattern support.

The most computationally intensive part of our method is to calculate distance $h(o, \mathcal{D})$ for each pattern P_i. To circumvent this problem, we introduce a distance matrix $\mathcal{M}$ that is spatially co-registered with the raster $\mathcal{R}$. The distance matrix $\mathcal{M}$ is pre-calculated to record the distance values of nearby pixels to the footprint of $\mathcal{D}$. Specifically, the cells in $\mathcal{M}$ that support at least one pattern ($o \in \cup F_i$) and are also located in $\mathcal{D}$ are set to the value of 1. For the remaining cells we calculate Hausdorff distance $h(o, \mathcal{D})$, using steps 4-7 of the algorithm mineCP. A gain in performance is achieved because pattern footprints often overlaps, and the distance is only calculated at most once for each pixel in $\mathcal{C}$. In steps 9-10, we calculate $|\mathcal{D}_+|^*$ and $|\mathcal{C}_+|^*$ (see Eqn. 5 and Eqn. 8) for each frequent pattern P_i, using the pre-calculated values in the distance matrix $\mathcal{M}$. In step 11, we calculate k_{P_i} to evaluate an aggregation a frequent pattern P_i. Finally, frequent patterns whose growth ratios are greater than the minimum growth-ratio threshold ρ are reported as controlling patterns.

Algorithm 1 mineCP: Mining Controlling Patterns

1: Mine frequent patterns in a transaction set $\mathcal{D}$ using a support threshold δ; output spatial footprint F_i of each frequent pattern P_i.
2: Initialize a distance matrix $\mathcal{M}$.
3: Assign every pixel in $\mathcal{R}$ to a correspondent cell in $\mathcal{M}$.
4: **for** each pixel $o \in (\cup F_i) \cap \mathcal{C}$ **do**
5: Calculate the distance $h(o, \mathcal{D})$.
6: Record the distance in corresponding cell of the distance matrix $\mathcal{M}$.
7: **end for**
8: **for** each frequent pattern P_i **do**
9: Calculate $|\mathcal{D}_+|^*$ and $|\mathcal{C}_+|^*$ with respect to P_i using the distance matrix $\mathcal{M}$.
10: Calculate growth ratio $CP_P^{\mathcal{D}^*} = \frac{|\mathcal{D}_+|^*/|\mathcal{D}|}{|\mathcal{C}_+|^*/|\mathcal{C}|}$.
11: Calculate k_{P_i}.
12: **end for**
13: **return** frequent patterns whose growth ratios $\geq \rho$.

4 Controlling Patterns of High Vegetation Density in the United States

In order to illustrate the working of our method and to demonstrate its utility, we applied it to a data set pertaining to a density of vegetation cover across the continental United States. The goal is to survey all combinations of controlling factors that are associated with high density of vegetation and to map their footprints. We address this goal by finding controlling patterns of high vegetation density. The class variable is 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 of vegetation density. The eleven plausible explanatory variables are summarized in Table 1; they 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, bulk density, permeability, porosity, 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 density, porosity, and permeability relate to the physical form of the soil. The dew temperature is an indicator of relative humidity. We made an attempt to use data pertaining to measurements performed at approximately the same time. These data sets are from different sources [19, 21, 25] and are available in different spatial resolutions. We have fused all the data sets to 11 co-registered latitude-longitude grids with a resolution of $0.5^o \times 0.5^o$. Each grid has 700 $\times$ 1253 pixels, of which 361,882 pixels (41.3%) have values for all the 11 variables. Note that this is an expository example intended only to illustrate the working of our method, because a more careful selection of a larger set of explanatory variables should be performed in ecological domain for an "industrial-strength" application.

All data sets are subjected to the categorization procedure described in Section 3.1 with six split points resulting in seven z-score bins $(-\infty, -2]$, $(-2, -1.5]$, $(-1.5, -0.5]$, $(-0.5, 0.5]$, $(0.5, 1.5]$, $(1.5, 2]$, and $(2, \infty)$, which are assigned categorical labels from 1 to 7, respectively. The NDVI data set is divided into two subsets, $\mathcal{D}$ with $cl = 1$, combining categories 6 and 7 of $(0.5, 1.5]$ and $(1.5, 2]$, and $\mathcal{C}$ with $cl = 0$, combining categories 1 to 5 of $(-\infty, -2]$, ..., $(0.5, 1.5]$. Thus, the vegetation density is arbitrarily defined as high when it is at least 1.5 $\times S_n$ higher than the median value. Fig. 1(b) shows the spatial distribution of $\mathcal{D}$. Each pixel carries a specific transaction consisting of the local values of explanatory variables and the class designation of vegetation density.

4.1 Results

We have conducted two numerical experiments. In the first primary experiment, we have employed the algorithm mineCP to identify controlling patterns of high vegetation density. In the second control experiment we have calculated controlling patterns using an algorithm that does not take advantage of a new definition of $\sup(P, \mathcal{D})$. In both experiments we used frequent pattern threshold $\delta = 0.2$ and the minimum growth-ratio threshold $\rho = 10.0$. Thus, we set up experiments to identify patterns whose footprints extend over at least 20% of the vegetation cover across the continental United States and which are at least 10 times more concentrated in the high vegetation density region. We also use $d = 1$ as a value of scale parameter in the footprint aggregation measure k_P (Eqn. 11). In the primary experiment, we use the influence function Ψ (Eqn. 7) with $\theta = 0.25$,

The primary experiment has identified 893 patterns frequent in $\mathcal{D}$, 780 of which have been determined to be controlling patterns. Fig. 3 shows the values of the growth ratio $CP_P^{\mathcal{D}^*}$ for all the 780 controlling patterns plotted against the values of k_P. The patterns are color-coded for the number of items present.

The blue dots indicate patterns consisting of up to 3 items; the green dots indicate patterns with 4 to 6 items; the red dots indicate patterns with 7 or more items. In general, the most interesting patterns are those indicated by the red dots and also located near the upper right corner of the graph. Such patterns are the most descriptive, highly specific to high vegetation density region, and they have highly aggregated footprints. Several immediate observations can be drawn from Fig. 3:

- All the controlling patterns have $k_P \geq 1$. This means that the complexes of controlling factors that are common in $\mathcal{D}$ are spatially aggregated.
- There appear to be some positive correlations between the values of the growth ratio $CP_P^{\mathcal{D}^*}$ and k_P. This indicates that patterns that are more indicative of high vegetation density are also more aggregated.
- Patterns with more attributes are more aggregated. More specific sets of controlling factors are restricted to more specific locations.

Tables 2 and 3 show the lists of the top 20 controlling patterns identified in the control and primary experiments, respectively. The 1^{st} column gives a pattern ID number; the 2^{nd} column shows the actual pattern; and the 3^{rd} column gives the number of features that match the pattern. The patterns are to be interpreted as described in Section 3.2. For example, pattern #56 (top emergent pattern on both lists) describes environment characterized by below average values (bin 3) of pH and above average values (bin 6, $(1.5, 2]$) of precipitation. Values of the remaining 9 explanatory variables vary from pixel to pixel within the footprint of this pattern. Thus, pattern #56 is not very specific as it involves only 2 out of 11 potential explanatory variables, but the conditions it describes exist in above 20% of the vegetation cover in the continental United States (frequent pattern) and almost nowhere outside the high vegetation density region (highly emergent pattern). Other emergent patterns are more descriptive; the most descriptive pattern in the top 20 controlling patterns (#832) involves 7 out of 11 explanatory variables.

The lists of controlling patterns stemming from primary and control experiments overlap, but they don't contain the same patterns; the patterns that don't occur in both lists are highlighted in Tables 2 and 3. Moreover, the order of patterns in the two lists is different. Thus, the modification of $\sup(P, \mathcal{D})$ leads to identification of different controlling patterns, or, at least, to different ordering of controlling patterns. We observe that pattern #12 is absent in the top 20 patterns identified in the primary experiment, though it is ranked as the 2^{nd} best in the control experiment. On the other hand, in the primary experiment, pattern #114 has improved its rank significantly ranking 2^{nd}, and new patterns, such as pattern #162, emerge in the top 20 list.

Does our new definition of $\sup(P, \mathcal{D})$ lead to "better" controlling patterns? In Section 3.3, we presented a heuristic argument for our modification, based on the notion that it offsets negative effects of data discretization. Do the results in Tables 2 and 3 support our argument? In order to address this problem quantitatively, we calculate "similarity" or degree of overlap between

the footprint of $\mathcal{D}$ (region of high vegetation density) and the footprints of the following patterns respectively,

- #12, 2^{nd} on the top 20 list resulting from the control experiment,
- #114, 2^{nd} on the top 20 list resulting from the primary experiment and 12^{th} on the top 20 list resulting from the control experiment,
- and #162, 8^{th} on the top 20 list resulting from the primary experiment.

The similarity is calculated using a measure based on mutual information between the two footprints [22]. The more the two footprints overlap the more one footprint determines the other footprint resulting in the larger value of mutual information. The mutual information-based similarity between the footprint of $\mathcal{D}$ and footprints of patterns #12, #114, #162 are 0.0188, 0.0214, and 0.0489, respectively. This shows that the footprints of patterns #114 and #162, whose ranks are increased while using the new definition of $\sup(P, \mathcal{D})$, match better with the region of high vegetation density than the footprint of pattern #12 whose rank decreases while using the new definition of $\sup(P, \mathcal{D})$. Thus, the new definition seems to promote patterns which offer a better spatial fit to the region of interest. Figures 4 (a-c) reiterate the same point in a graphical fashion; compared with the footprint of pattern #12, footprints of patterns #114 and #162 align better with the footprint of high vegetation density region.

To further demonstrate the advantage of using a modified definition of $\sup(P, \mathcal{D})$, we compare the unions of footprints of all 20 patterns stemming from the primary and control experiments, respectively. Figures 5(a-b) show such comparison; careful examination reveals that the union of footprints shown in Fig. 5(a) is more "land filling," exactly the effect we intended to achieve by our modification. The mutual information-based similarity between the footprint of $\mathcal{D}$ and the unions of footprints of the 20 top pattern calculated in the primary and control experiments is 0.0649 and 0.0563, respectively.

It is instructive to examine the values of explanatory variables as they appear as items in the top 20 patterns in Table 3. Interestingly, top controlling patterns contain only a limited range of values for each variable. For example, in the patterns that contain average annual maximum temperature (tmax) only values of tmax=5 (0.5 to 1.5 $\times S_n$ above the median value of tmax) are present. Similarly, top controlling patterns contain only the values pH=3 (0.5 to 1.5 $\times S_n$ below the median value of pH). In general the range is at most two consecutive bins. This indicate that environmental conditions that support high vegetation density are restricted to rather narrow values of explanatory variables, although, within these limits, the specific conditions may vary between different geographical locations.

5 Conclusions

In this paper, we have presented a methodology for surveying and mapping relationship between predictors and response in geospatial data sets. In particular, our method reveals diversity of predictors associated with the same

response. In departure from a previous attempt to address this problem [29], we base our method on the association analysis technique of mining for emerging patterns which offers a principled approach to the problem and also has several potential advantages. First, the model (at its core) is conceptually strong and transparent—the only numerical operation is pixel counting. This intrinsic simplicity of association analysis assures that all nonlinearities in the system are taken fully into account. Second, the method yields itself into modification that accommodates the spatial character of the data. We show how to modify the standard association analysis methodology to accommodate spatial data sets. Evaluation of the vegetation data set confirmed that those modifications lead to somewhat improved result over an application of unmodified method.

Our approach, in its present form, has some shortcomings that need to be addressed by further research. One issue is the need for data categorization. This issue enters only in the context of dividing a response variable into the two classes. In the present paper we addressed this issue by modifying the definition of pattern support so it takes into consideration spatial proximity. Future research would concentrate on further improvement of the definition of pattern support, so it takes into consideration not only spatial proximity but also feature similarity. Another issue is the difficulty in interpreting the outcome of our method. This is a common problem with the association analysis; it produces thorough but large output that frequently requires a summarization technique in order to presented the results in a comprehensive fashion. In our application to the vegetation data set, we have identified 780 controlling patterns. Each controlling pattern represents a nugget of knowledge about the local combination of predictors associated with the phenomenon. However, this association is not exclusive, other patterns may also cover the same location. In general, footprints associated with controlling patterns are not mutually exclusive and exhaustive, instead they overlap and their union does not necessarily cover the entire extent of the phenomena. In Section 4, we have showed top 20 controlling patterns as well as footprints for few patterns, more extensive visualization would take significantly more space. In order for the results of our method to be more effectively presented a companion method of pattern summarization needs to be developed. We have already taken the first step into this direction by investigating a possibility to define a similarity measure between the patterns [8]. Such similarity could be used to cluster the patterns into a smaller number of "super-patterns"—a sets of patterns having similar meaning and sharing common spatial extent [8]. Super-patterns would provide a more concise means to visualize our output and they would make possible a direct comparison of our results with the results of the method based on the regression tree [29].

6 Acknowledgments

The work is supported in part by the National Science Foundation under Grant IIS-0812271. 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. This is LPI Contribution No.xxx.

References

1. R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, pages 207–216, Washington, D.C., 26–28 1993.
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. N. A. Cressie. *Statistics for Spatial Data*. Wiley, 1993.
7. W. Ding, T. F. Stepinski, R. Parmar, D. Jiang, and C. F. Eick. Discovery of feature-based hot spots using supervised clustering. *Computers and Geosciences*, in press, 2009.
8. W. Ding, T. F. Stepinski, and J. Salazar. Discovery of geospatial discriminating patterns from remote sensing datasets. In *SIAM International Conference on Data Mining (SDM), Nevada, April 2009*, 2009.
9. 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.
10. T. Korkalainen and A. Lauren. Using phytogeomorphology, cartography and GIS to explain forest site productivity expressed as tree height in southern and central finland. *Geomorphology*, 74:271–284, 2006.
11. D. R. Larsen and P. L. Speckman. Multivariate regression trees for analysis of abundance data. *Biometrics*, 60 (2):543–549, 2004.
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 L. Wong. Structural geography of the space of emerging patterns. *Intelligent Data Analysis*, 9(6):567–588, 2005.
14. 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.
15. T. Liaghati, M. Preda, and M. Cox. Heavy metal distribution and controlling factors within coastal plain sediments, Bells Creek catchment, southeast Queensland, Australia. *Environment International*, 29:935–948, 2003.
16. D. B. Lobell, J. I. Ortiz-Monasterio, G. P. Asner, R. L. Naylor, and W. P. Falcon. Combining field surveys, remote sensing, and regression trees to understand yield variations in an irrigated wheat landscape. *Agron. J.*, 97:241–249, 2005.
17. J. Munkres. *Topology*. Prentice Hall, 2nd edition, 1999.
18. A. Navas and J. Machn. Spatial distribution of heavy metals and arsenic in soils of Aragn (northeast Spain): controlling factors and environmental implications. *Applied Geochemistry*, 17:961–973, 2002.
19. ORNL. Oak Ridge National Laboratory distributed active archive center data holdings, 2009.

20. 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.
21. PRISM. PRISM (parameter-elevation regressions on independent slopes model) climate mapping system products matrix, 2009.
22. T. K. Remmel and F. Csillag. Mutual information spectra for comparing categorical maps. *International Journal of Remote Sensing*, 27:1425–1452, 2006.
23. J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. American Stat. Association*, 88:1273–1283, 1993.
24. S. Rusjan and M. Mikos. Assessment of hydrological and seasonal controls over the nitrate flushing from a forested watershed using a data mining technique. *Hydrol. Earth Syst. Sci.*, 12:645–656, 2008.
25. Seamless. National Map Seamless Server, 2009.
26. A. Steegen, G. Govers, I. Takkena, J. Nachtergaelea, J. Poesena, and R. Merckxb. Factors controlling sediment and phosphorus export from two belgian agricultural catchments. *Journal of Environmental Quality*, 30:1249–1258, 2001.
27. T. Stepinski, W. Ding, and C. Eick. Discovering controlling factors of geospatial variables. In *the 16th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (ACM GIS 2008)*, pages 1–4, Irvine, CA, USA, November 2008.
28. X. Wang and Y. Qin. Spatial distribution of metals in urban topsoils of xuzhou (china): controlling factors and environmental implications. *Environmental Geology*, 49(6):905–914, 2005.
29. D. White and J. C. Sifneos. Regression tree cartography. *J. Computational and Graphical Statistics*, 11 (3):600–614, 2002.
30. I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

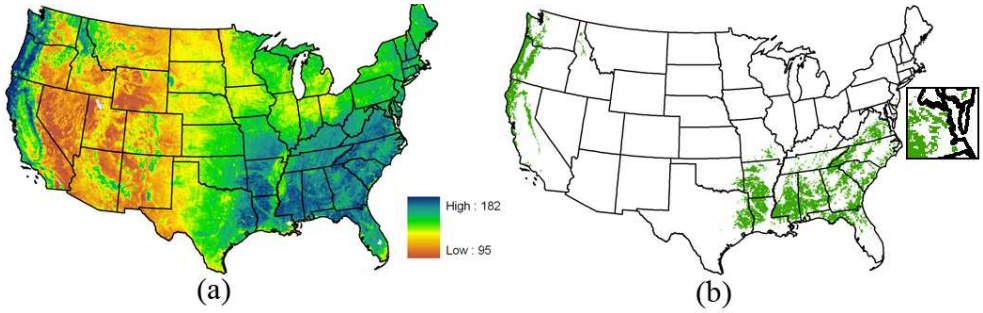

Fig. 1 (a) Map of vegetation density in the United States. The value of Normalized Difference Vegetative Index (NDVI) serves as proxy for vegetation density. (b) Footprint of high vegetation density region $\mathcal{D}$ defined by the two highest categorical bins of class variable is shown in green. A zoomed-in window centered on the states Virginia and Maryland shows sharp boundaries of high vegetation density footprint in details.

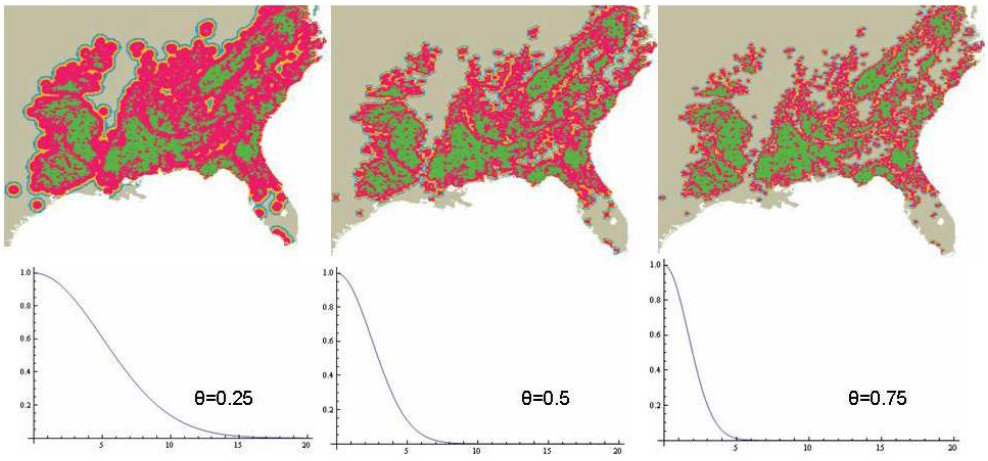

Fig. 2 Illustrating a new method of calculating pattern support by including fractional support from nearby pixels. Original footprint of high vegetation density region is shown in green (weight= 1), pixels contributing weights in the range $(1, 0.5)$ are shown in red, those contributing weights in the range $(0.5, 0.25)$ are shown in yellow, in the range $(0.25, 0.1)$ are shown in blue, and those contributing less than 0.1 are shown in gray. The result depends on the value of θ: 0.25 (left), 0.5(middle), and 0.75 (right).

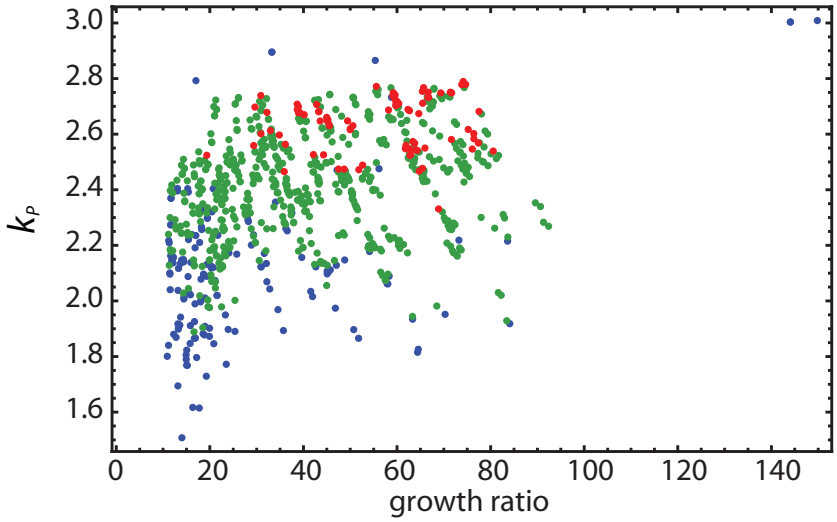

Fig. 3 Properties of 780 controlling patterns of high vegetation density in the United States. Colors: blue - patterns with 1-3 items, green - patterns with 4-6 items, and red - patterns with ≥ 7 items.

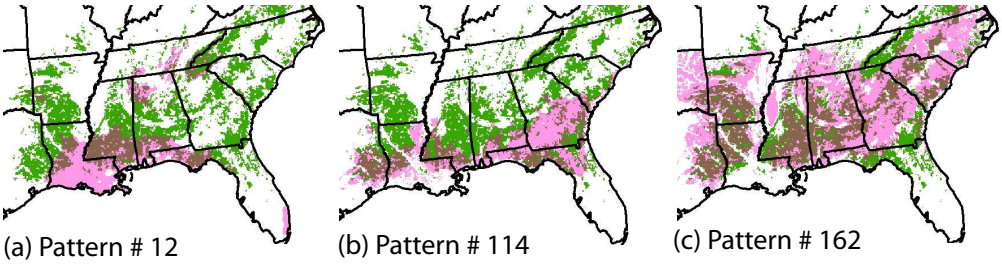

Fig. 4 (a-c) Footprints of patterns #12, #114, #162. Colors: green - high vegetation density region, pink - footprints of patterns, dark brown - overlays between pattern footprint and high vegetation density region.

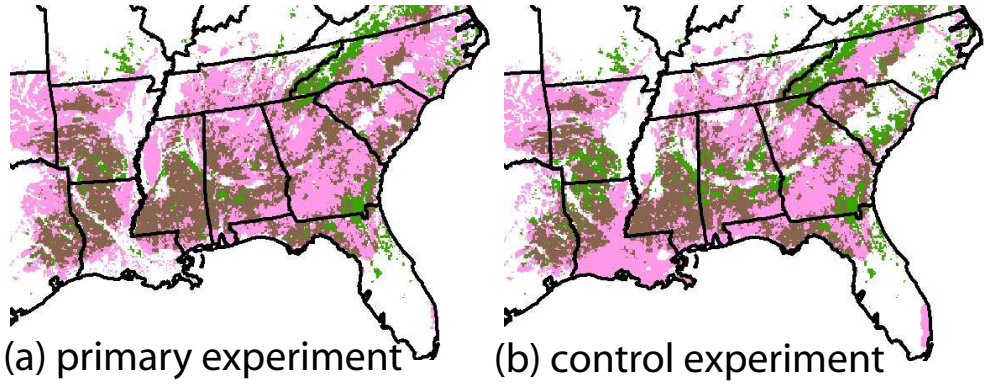

Fig. 5 Union of footprints for the top 20 patterns stemming from the primary experiment (a) and control experiment (b). Colors: green - high vegetation density region, pink - union of footprints, dark brown - overlays between the union of footprints and high vegetation density region.

Table 1 Data sets of explanatory variables used in the case study

Variable Abbreviation	Short Description
1. awc	Available water capacity (2000) ORNL for biogeochemical and ecological data
2. bd	Soil bulk density (2000) ORNL for biogeochemical and ecological data
3. dew	Average dew point temperature (annual 2005) PRISM climate mapping system
4. elev	Elevation USGS National Map Seamless Server
5. perm	Soil permeability (2000) ORNL for biogeochemical and ecological data
6. ph	Soil pH (2000) ORNL for biogeochemical and ecological data
7. poros	Soil porosity (2000) ORNL for biogeochemical and ecological data
8. ppt	Average annual precipitation (1971 - 2001) PRISM climate mapping system
9. tmax	Average annual maximum temperature (1971 - 2001) PRISM climate mapping system
10. tmin	Average annual minimum temperature (1971 - 2001) PRISM climate mapping system
11. aveveg	Vegetation growth average (annual 2005) USGS National Map Seamless Server

Table 2 Top 20 controlling patterns found in the control experiment

Pattern ID	Patterns	# of Variables
56	ph=3, ppt=6	2
12	ppt=6	1
312	awc=4, perm=4, ph=3, tmax=5	4
586	awc=4, perm=4, ph=3, tmax=5, tmin=5	5
555	awc=4, elev=3, perm=4, ph=3, tmax=5	5
780	awc=4, elev=3, perm=4, ph=3, tmax=5, tmin=5	6
135	perm=4, ph=3, tmax=5	3
337	perm=4, ph=3, tmax=5, tmin=5	4
314	elev=3, perm=4, ph=3, tmax=5	4
588	elev=3, perm=4, ph=3, tmax=5, tmin=5	5
114	dew=6, elev=3, ph=3	3
34	dew=6, ph=3	2
318	awc=4, perm=4, ph=3, tmin=5	4
568	awc=4, elev=3, perm=4, ph=3, tmin=5	5
530	awc=4, bd=4, perm=4, ph=3, tmax=5	5
695	awc=4, dew=5, perm=4, ph=3, tmax=5, tmin=5	6
765	awc=4, bd=4, perm=4, ph=3, tmax=5, tmin=5	6
549	awc=4, perm=4, ph=3, poros=5, tmax=5	5
441	awc=4, dew=5, perm=4, ph=3, tmax=5	5
776	awc=4, perm=4, ph=3, poros=4, tmax=5, tmin=5	6

Table 3 Top 20 controlling patterns found in the primary experiment

Pattern ID	Patterns	# of Variables
56	ph=3, ppt=6	2
114	dew=6, elev=3, ph=3	3
34	dew=6, ph=3	2
312	awc=4, perm=4, ph=3, tmax=5	4
555	awc=4, elev=3, perm=4, ph=3, tmax=5	5
586	awc=4, perm=4, ph=3, tmax=5, tmin=5	5
780	awc=4, elev=3, perm=4, ph=3, tmax=5, tmin=5	6
162	awc=4, ph=3, tmax=5	3
314	elev=3, perm=4, ph=3, tmax=5	4
135	perm=4, ph=3, tmax=5	3
360	awc=4, elev=3, ph=3, tmax=5	4
337	perm=4, ph=3, tmax=5, tmin=5	4
588	elev=3, perm=4, ph=3, tmax=5, tmin=5	5
393	awc=4, ph=3, tmax=5, tmin=5	4
695	awc=4, dew=5, perm=4, ph=3, tmax=5, tmin=5	6
639	awc=4, elev=3, ph=3, tmax=5, tmin=5	5
441	awc=4, dew=5, perm=4, ph=3, tmax=5	5
318	awc=4, perm=4, ph=3, tmin=5	4
832	awc=4, dew=5, elev=3, perm=4, ph=3, tmax=5, tmin=5	7
671	awc=4, dew=5, elev=3, perm=4, ph=3, tmax=5	6