Discovering Controlling Factors of Geospatial Variables

Tomasz Stepinski
Lunar and Planetary Institute
tom@lpi.usra.edu

Wei Ding
UMass-Boston
ding@cs.umb.edu

Christoph F. Eick
University of Houston
ceick@uh.edu

ABSTRACT

Efficient means of determining factors controlling spatial distribution of an environmental class variable are of significant interest in Earth science. In this paper, we present a method for automated discovery of controlling factors by mining for emerging patterns in a database constructed from the fusion of several explanatory datasets. We introduce a new definition of pattern support to account for spatial character of the data and systematically evaluate the effectiveness of our technique using a real-world application pertaining to density of vegetation cover. Experimental results show that our method can successfully identify controlling factors for the presence of high vegetation cover.

Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Pattern analysis*; I.5.4 [**Pattern Recognition**]: Applications—*Geoscience*

Keywords

Controlling factors, Emerging patterns, Spatial data mining

1. INTRODUCTION

Environmental variables are highly coupled through a complex chain of interactions resulting in their mutual interdependability. Efficient means of determining controlling factors which are responsible for spatial distribution of a selected class variable are of significant interest to domain experts. In order to understand the properties of a class variable, experts attempt to identify dominant *controlling factors*: explanatory variables that are predominantly responsible for spatial distribution of the class variable. For example, the density of vegetation (class variable) shows large regional variations due to changes in environmental variables such as various climate indicators and soil properties. Understanding the dominant factors behind geographical extent of high vegetation regions is of great interest from a scientific as well as a practical point of view, such as the impact of global warming on vegetation cover.

Currently, such problems are addressed by painstaking manual analysis, but with an increasing access to data products describing terrain topography, climate, vegetation cover, lithology, soil properties, human impact, etc., a more comprehensive methodology, one that would distill all dependencies in the data and present an expert with a relevant summary needs to be developed. In this paper, we propose a method for machine-aided identification of dominant controlling factors for a geospatial class variable. The method extends the standard technique of mining for emerging patterns (EPs) [2, 3] to geospatial domain.

Identifying EPs in spatial datasets presents challenges. Firstly, all databases of interest are spatial databases; this requires modification of standard EPs methods. Secondly, relevant geospatial databases 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 boundaries between different regions (for example, regions of high vegetation and low vegetation) where in reality the regions are not mutually exclusive, but instead they gradually transit into each other. We address these challenges by generalizing the definition of pattern support. The new definition takes into account spatial autocorrelation - a common feature in spatially extended variables. The overall goal of the method described in this paper is to provide means for discovering a set of dominant controlling factors that provides comprehensive and systematic knowledge for building empirical models of a chosen phenomena.

We illustrate our method by applying it to a real-world case study of finding factors exerting control over the density of vegetation cover in the continental United States. Continuing with the example given above, a territory of the United States can be divided into two sets based on the values of a class variable: high vegetation density region, and not-high vegetation density region. The patterns to be considered are the collections of the values of environmental explanatory variables. The goal is to identify patterns that are relatively frequent in the high vegetation region but rare in the low vegetation region. Examination of patterns related to the high vegetation provides a summary of data dependencies that helps to develop a better empirical models of the vegetation growth.

2. METHODOLOGY

2.1 Data Preprocessing

ACM GIS '08, November 5-7, 2008. Irvine, CA, USA

The numerical values of different explanatory variables are distributed differently. 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 μ, a robust estimator is the trimmed mean calculated by discarding a certain percentage of the lowest and the highest values. We use the median - a particular example of the trimmed mean - to estimate μ. For σ, we use the S_n estimator introduced by [9]: $S_n = c\ med_i\{med_j|x_i - x_j|\}$, where c is a constant and its value is 1.1926, and 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. S_n gives a robust estimate of σ regardless whether the distribution is symmetric or asymmetric.

In order to categorize the data, we use the values of μ and σ to transform the continuously distributed variables into their *modified z-scores*. 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. For simplicity, we use the term z-score for the modified z-score in the rest of the paper. Two different variables with the same z-score are "equal" in the sense that both are deviated by the same relative amount from their medians. The actual discretization is achieved by assigning the z-scores into n bins using $n-1$ split points. This transforms all real-valued datasets into categorical datasets with a common range.

2.2 Problem Definition

The fusion of all the datasets relevant to a given task results in a geospatial dataset $\mathcal{R}$. The objects in $\mathcal{R}$ are tuples having the following form $r = \{x, y; a_1, a_2, ..., a_m; cl\}$, where the first two entries are spatial coordinates, the next m entries are categorical values of m explanatory variables that can potentially exert control over the class variable, and the last entry 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 as a transaction $\{a_1, a_2, ..., a_m; cl\}$. All transactions are classified into two mutually exclusive and exhaustive sets: dataset $\mathcal{D}$ grouping transactions with $cl = 1$ and dataset $\mathcal{C}$ grouping transactions with $cl = 0$. A pattern (itemset) is a set of items contained in a transaction. A transaction supports the pattern P if it has such values at the indicated positions. The footprint of the pattern P is a projection of the objects that support P into the spatial reference system of the dataset $\mathcal{R}$.

We defined a controlling pattern as:

Definition 1. A **controlling pattern (CP)** P in $\mathcal{D}$ is an itemset such that its growth ratio $CP_P^{\mathcal{D}} = \frac{\sup(P,\mathcal{D})}{\sup(P,\mathcal{C})} \geq \rho$, where ρ is a user-defined minimum growth-ratio threshold. $\sup(P, \mathcal{D})$ and $\sup(P, \mathcal{C})$ are the support of a pattern P in $\mathcal{D}$ and $\mathcal{C}$, respectively.

We define such itemsets as controlling patterns, because they correspond to particular values of certain explanatory variables that happen to be associated disproportional with $cl = 1$ objects. It is therefore expected that they constitute controlling factors for the distribution of $cl = 1$ objects.

2.3 Calculating Pattern Support

Let's $\mathcal{F}$ be a set of transactions (objects) that support a given pattern P, and $\mathcal{G}$ be a set of transactions where P is absent. We define the following sets: $\mathcal{D}_+ = \mathcal{D} \cap \mathcal{F}$, $\mathcal{D}_- = \mathcal{D} \cap \mathcal{G}$, $\mathcal{C}_+ = \mathcal{C} \cap \mathcal{F}$, $\mathcal{C}_- = \mathcal{C} \cap \mathcal{G}$. The support of P in datasets $\mathcal{D}$ and $\mathcal{C}$ is defined as: $\sup(P, \mathcal{D}) = \frac{|\mathcal{D}_+|}{|\mathcal{D}|}$, $\sup(P, \mathcal{C}) = \frac{|\mathcal{C}_+|}{|\mathcal{C}|}$. Thus,

$$CP_P^{\mathcal{D}} = \frac{\sup(P,\mathcal{D})}{\sup(P,\mathcal{C})} = \frac{|\mathcal{D}_+|/|\mathcal{D}|}{|\mathcal{C}_+|/|\mathcal{C}|} \tag{1}$$

where, $|\ |$ denotes the number of elements in a set. Notice that $|\mathcal{D}_+| + |\mathcal{D}_-| + |\mathcal{C}_+| + |\mathcal{C}_-| = |\mathcal{R}|$. Discovering CP is a matter of evaluating $CP_P^{\mathcal{D}}$, given by Eqn. 1 for a set of patterns and selecting those patterns that have $CP_P^{\mathcal{D}} \geq \rho$.

The set $\mathcal{D}$ may be defined by an arbitrary threshold. Because of spatial continuity of geospatial data many objects (pixels) nearby the footprint of $\mathcal{D}$ are expected to have vegetation values, that although not high enough to be in the two highest bins, are nevertheless high enough to be in the upper range of the highest categorical bin not included in the definition of "high" vegetation. In order to take such spatial continuity into consideration, we introduce a new definition of pattern support so it also accounts for tuples located nearby $\mathcal{D}$.

Specifically, we propose a following modification of $|\mathcal{D}_+|$ that we denote by $|\mathcal{D}_+|^*$: $|\mathcal{D}_+|^* = \sum_{o \in \mathcal{F}} w(o, \mathcal{D})$, where o is a spatial object and $w(o, \mathcal{D})$ is a weight determined on the basis of spatial proximity of this object to the set $\mathcal{D}$. The weight is calculated using the following formula:

$$w(o, \mathcal{D}) = \begin{cases} 1 & o \in \mathcal{D} \\ \Psi(h(o, \mathcal{D})) & o \in \mathcal{C} \end{cases} \tag{2}$$

The influence function Ψ determines the weight for objects outside of the footprint. In principle, an influence function can be an arbitrary function. In this paper, we use a half normal influence function, which is a normal distribution with mean 0:

$$\Psi(\xi) = exp(\frac{-\theta^2 \xi^2}{\pi}) \tag{3}$$

where θ ($\theta \in [0, \infty)$) is a free parameter and π is the mathematical π. The function Ψ determines the weight for objects outside of the footprint: the farther a spatial object o is from the dataset $\mathcal{D}$, the less weight is for the object. Nearby pixels have weights decreasing with increasing distance from the region. The less the θ value, the more surrounding objects will be taken into account.

The function $h(o, \mathcal{D})$ used by the influence function Ψ is a special case of Hausdorff distance [4] in the spatial domain, where the function measures the minimum distance between o to the nearest high-vegetation object in $\mathcal{D}$.

Thus, in the proposed modification, the support of the pattern P "in" $\mathcal{D}$ is increased (because $|\mathcal{D}_+|* > |\mathcal{D}_+|$), if a significant number of objects close to the footprint of $\mathcal{D}$ conform to this pattern. Simultaneously, in such a situation, the support P in $\mathcal{C}$ must be decreased by the same amount. This is achieved by defining $|\mathcal{C}_+|^* = \sum_{o \in \mathcal{F}} [1 - w(o, \mathcal{D})]$.

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

$$CP_P^{\mathcal{D}*} = \frac{|\mathcal{D}_+|^*/|\mathcal{D}|}{|\mathcal{C}_+|^*/|\mathcal{C}|} \tag{4}$$

Table 1: 11 geospatial datasets used in our case study

Variable	Short Description
awc	Available water capacity [6]
bd	Soil bulk density [6]
dew	Average dew point temperature [7]
elev	Elevation [5]
perm	Soil permeability[6]
ph	Soil pH [6]
poros	Soil porosity [6]
ppt	Average annual precipitation [7]
tmax	Average annual maximum temperature [7]
tmin	Average annual minimum temperature [7]
aveveg	Vegetation growth average [5]

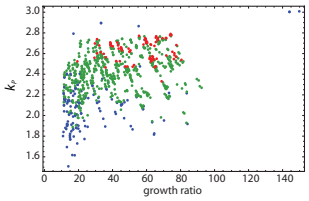

Figure 1: Diagrams describing 780 controlling patterns using the new definition of pattern support. Blue: patterns with 1-3 items, green: 4-6 items, and red: $\geq$ 7 items.

Note that in Eqn. 4 we do not modify the values of $|\mathcal{D}|$ and $|\mathcal{C}|$, and the total number of spatial objects remains the same: $|\mathcal{D}_+|^* + |\mathcal{D}_-| + |\mathcal{C}_+|^* + |\mathcal{C}_-| = |\mathcal{R}|$.

2.4 Measuring Spatial Aggregation of Pattern Footprints

In addition to the growth ratio $CP_P^{\mathcal{D}^*}$, we propose that in the case of spatial datasets it is also important to measure the degree of aggregation of the footprint of a pattern P. The patterns with more aggregated footprints are more interesting to domain experts as they more likely point to real controlling factors. We use Ripley's K function, a statistical method frequently applied to point pattern analysis, to quantify such aggregation. Following the concept of nearest neighbor analysis, Riley's K function quantifies the spatial pattern intensity of points in a circular search window. Without considering edge effects, Ripley's K function can be estimated as $\hat{K}(d) = \frac{F}{N^2}\sum_{i=1}^{N}\sum_{j=1,j\neq i}^{N} I_d(d_{ij})$ [1], where N is the number of pixels (points) in the footprint of the pattern P, d_{ij} is the distance between the i^{th} and j^{th} points, $I_d(d_{ij})$ 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 the pattern 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 $K_o(d) = \pi d^2$,

$$k_P = \sqrt{\frac{\hat{K}(d)}{K_o(d)}} = \frac{\sqrt{\frac{\hat{K}(d)}{\pi}}}{d} \tag{5}$$

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.

Table 2: Top 20 patterns

ID	Patterns Using the Traditional Definition
56	ph=3 ppt=6
12	ppt=6
312	awc=4 perm=4 ph=3 tmax=5
586	awc=4 perm=4 ph=3 tmax=5 tmin=5
555	awc=4 elev=3 perm=4 ph=3 tmax=5
780	awc=4 elev=3 perm=4 ph=3 tmax=5 tmin=5
135	perm=4 ph=3 tmax=5
337	perm=4 ph=3 tmax=5 tmin=5
314	elev=3 perm=4 ph=3 tmax=5
588	elev=3 perm=4 ph=3 tmax=5 tmin=5
114	dew=6 elev=3 ph=3
34	dew=6 ph=3
318	awc=4 perm=4 ph=3 tmin=5
568	awc=4 elev=3 perm=4 ph=3 tmin=5
530	awc=4 bd=4 perm=4 ph=3 tmax=5
695	awc=4 dew=5 perm=4 ph=3 tmax=5 tmin=5
765	awc=4 bd=4 perm=4 ph=3 tmax=5 tmin=5
549	awc=4 perm=4 ph=3 poros=5 tmax=5
441	awc=4 dew=5 perm=4 ph=3 tmax=5
776	awc=4 perm=4 ph=3 poros=4 tmax=5 tmin=5

ID	Patterns Using the New Definition
56	ph=3 ppt=6
114	dew=6 elev=3 ph=3
34	dew=6 ph=3
312	awc=4 perm=4 ph=3 tmax=5
555	awc=4 elev=3 perm=4 ph=3 tmax=5
586	awc=4 perm=4 ph=3 tmax=5 tmin=5
780	awc=4 elev=3 perm=4 ph=3 tmax=5 tmin=5
162	awc=4 ph=3 tmax=5
314	elev=3 perm=4 ph=3 tmax=5
135	perm=4 ph=3 tmax=5
360	awc=4 elev=3 ph=3 tmax=5
337	perm=4 ph=3 tmax=5 tmin=5
588	elev=3 perm=4 ph=3 tmax=5 tmin=5
393	awc=4 ph=3 tmax=5 tmin=5
695	awc=4 dew=5 perm=4 ph=3 tmax=5 tmin=5
639	awc=4 elev=3 ph=3 tmax=5 tmin=5
441	awc=4 dew=5 perm=4 ph=3 tmax=5
318	awc=4 perm=4 ph=3 tmin=5
832	awc=4 dew=5 elev=3 perm=4 ph=3 tmax=5 tmin=5
671	awc=4 dew=5 elev=3 perm=4 ph=3 tmax=5

3. CASE STUDY: CONTROLLING PATTERNS FOR HIGH VEGETATION COVER

In order to show the utility of our method for discovering controlling patterns, we have constructed a fusion of several geospatial datasets that pertain to the distribution of topography, climate, and soil properties across the continental United States. The purpose is to identify dominant factors responsible for the regions of high vegetation cover. The datasets are summarized in Table 1. The 10 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, bulk density, permeability, porosity, and soil pH), and topography-related (elevation). We have fused all the datasets to 11 co-registered latitude-longitude grids with a resolution of $0.5^o \times 0.5^o$. Each grid has 700×1253 pixels, of which 361,882 pixels (41.3%) have values for all the 11 variables.

All the co-located datasets are subjected to a categorization procedure 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 vegetation density dataset is divided into two subsets, $\mathcal{D}$ with $cl = 1$ (combined categories 6 and 7) and $\mathcal{C}$ with $cl = 0$ (combined categories 1 to 5).

3.1 Experimental Results

Setting the support threshold for a frequent pattern in $\mathcal{D}$ to $\delta = 0.2$ and the minimum growth-ratio threshold $\rho =$

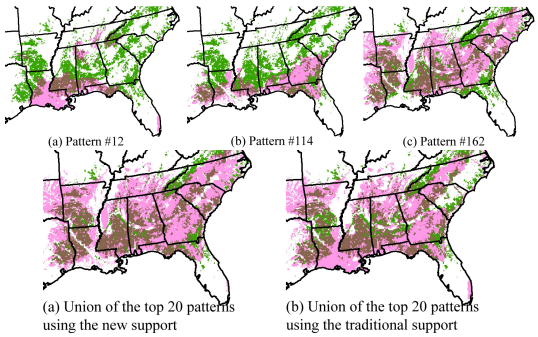

Figure 2: Top row: Footprints of patterns #12, #114, #162. Bottom row: Union of footprints for the top 20 patterns. Green: high vegetation cover, pink: footprints of patterns, dark brown: overlays between the footprints and high vegetation cover.

10.0, we have found 780 controlling patterns. Fig. 1 [1]shows the values of the growth ratio $CP_P^{\mathcal{D}^*}$ for all the 780 controlling patterns plotted against the values of k_P. We use the particular choice of the influence function Ψ of Eqn. 3 using $\theta = 0.25$, and $d = 1$ for the k_P in Eqn. 5. Several immediate observations can be drawn from Fig. 1: (1) all the controlling patterns have $k_P \geq 1$. This means that the complexes of controlling factors that are common in $\mathcal{D}$ are also spatially focused on $\mathcal{D}$ as they are less common in $\mathcal{C}$. (2) There appear to be some positive correlations between values of the growth ratio $CP_P^{\mathcal{D}^*}$ and k_P. This indicates that patterns that are more indicative of high vegetation are also more aggregated. And (3) patterns with more attributes are more aggregated. More specific sets of controlling factors are restricted to more specific locations.

Table 2 shows the top 20 patterns using the traditional definition of pattern support using Eqn. 1 and the new definition of pattern support using Eqn. 4. The 1st column gives a pattern ID number; the 2nd column shows the actual pattern; and the 3rd column gives the number of items that match the pattern. Modification of the support definition causes different values for growth ratio resulting in different ordering of controlling factors. The patterns that do not occur in both top 20 lists are highlighted in Table 2. We observe that pattern #12 is absent in the top 20 patterns using the new definition of pattern support, though it is ranked as the 2nd best using the traditional definition. On the other hand, pattern #114 has improved its rank significantly ranking 2nd, and new patterns such as pattern #162, emerge in the top 20 list when using the new definition of support. We then use the mutual information measure (MI) to check whether the footprints of patterns #114 and #162 match better with the footprints of high vegetation. Mutual information has been introduced in [8] for 2D image matching. The method tries to find the mutual information between the two images using $MI(X,Y) = H(Y) - H(Y|X)$, where X and Y are images and function $H()$ is the entropy function. $MI(X,Y)$ is a measure of the reduction of the entropy of Y given X. In principle, if two images are correctly matched, one image can give information about the other, thus their mutual information is high. The mutual information between the footprints of high vegetation cover and patterns #12, #114, #162 are 0.0188, 0.0214, and 0.0489, respectively. The result shows that patterns #114 and #162, whose ranks are increased using the new support, match better with the regions of high vegetation. For illustration, Figures 2 top row (a-c) depict the spatial footprints of the three patterns #12, #114, #162. Compared with pattern #12 in Fig. 2 top row (a), Patterns #114 in Fig. 2 top row (b) and #162 in Fig. 2 top row (c) align better with the footprint of high vegetation. The results indicate that using the new definition of pattern support, we can identify better controlling patterns than would be possible using the traditional definition.

Figures 2 bottom row (a-b) show a comparison in the spatial coverage of the top 20 patterns in Table 2. The area shown is the southern portion of the continental United States. There are important differences between the two footprints, with the coverage shown in Fig. 2 bottom row (a) being more "land filling," exactly the effect expected to be revealed by our method. To compare the quality of the top 20 patterns, we calculate MI between the footprint of high vegetation cover and the union of top 20 patterns using the traditional and new definition of support, respectively. We obtain 0.0649 for the footprints using new pattern support and 0.0563 for those using the traditional pattern support. The top 20 patterns identified using our improved method are more restricted to high vegetation regions, thus they can better control the presence of high vegetation cover.

To summarize, we have formulated the problem of mining controlling factors and proposed a new definition of pattern support in the domain of geoscience based on the concept of emerging patterns. In the case study pertaining to vegetation cover across United States our method identifies dominant patterns that control high vegetation density.

[1]Figs 1 and 2 are better viewed in color, view an electronic color version of the paper at http://www.cs.umb.edu/∼ding.

4. ACKNOWLEDGEMENT

We acknowledge support by NSF (0430208) and NASA (NNG06GE57G). A portion of this research was conducted at the Lunar and Planetary Institute under contract CAN-NCC5-679 with NASA. This is LPI Contribution No. 1434.

5. REFERENCES

[1] N. A. Cressie. *Statistics for Spatial Data*. Wiley, 1993.
[2] G. Dong and J. Li. Efficient mining of emerging patterns: discovering trends and differences. In *Proc. of the 5th ACM SIGKDD intl. conf. on knowledge discovery and data mining*, pages 43–52, California, United States, 1999.
[3] J. Li and Q. Yang. Strong compound-risk factors: Efficient discovery through emerging patterns and contrast sets. *IEEE Trans. on Information Technology in Biomedicine*, 11:544–552, 2007.
[4] J. Munkres. *Topology*. Prentice Hall, 2nd edition, 1999.
[5] National Map Seamless Server.
[6] Oak Ridge National Laboratory Distributed Active Archive Center Data Holdings.
[7] PRISM (Parameter-elevation Regressions on Independent Slopes Model) Climate Mapping System Products Matrix.
[8] T. K. Remmel and F. Csillag. Mutual information spectra for comparing categorical maps. *intl. J. of Remote Sensing*, 27:1425–1452, 2006.
[9] J. Rousseeuw and C. Croux. Alternatives to the median absolute deviation. *J. American Stat. Association*, 88:1273–1283, 1993.