Environ Ecol Stat (2006) 13:365–377
DOI 10.1007/s10651-006-0017-5

ORIGINAL ARTICLE

Spatially constrained clustering and upper level set scan hotspot detection in surveillance geoinformatics

G.P. Patil · R. Modarres · W.L. Myers · P. Patankar

Received: September 2005 / Revised: February 2006

Abstract We discuss upper level set (ULS) scan as a type of spatially constrained clustering in relation to two ways of imposing the spatial constraint, retrospectively versus progressively. We show that ULS scan produces the same results both ways; whereas two popular clustering techniques, single-linkage and K-means, can yield different results when spatial constraints are imposed retrospectively versus progressively. The ULS scan approach examines spatially connected components of a tessellation as a threshold is moved from the highest level (value) in the data to the lowest level. When the variable of interest on the tessellation is a rate of incidence, then a significance test is available based on binomial or Poisson null models and Monte Carlo techniques. This is a common context for detecting hotspots of diseases in epidemiological work. We also discuss an approach for extending the univariate methodology to accommodate multivariate contexts.

This material is based upon work supported by (i) the National Science Foundation under Grant No. 0307010, (ii) the United States Environmental Protection Agency under Grant No. CR-83059301 and (iii) the Pennsylvania Department of Health using Tobacco Settlement Funds under Grant No. ME 01324. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the agencies.

G. P. Patil (✉)
Department of Statistics, The Pennsylvania State University, University Park, PA, USA
e-mail: ees@stat.psu.edu

R. Modarres
Department of Statistics, George Washington University, Washington, USA

W. L. Myers
School of Forest Resources and Office of Remote Sensing of Earth Resources, Penn State Institutes of Environment, The Pennsylvania State
University, University Park, PA, USA

P. Patankar
Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA, USA

Keywords Hotspot detection · Multivariate scan by spatial coincidence · Spatially constrained clustering · Surveillance geoinformatics · Upper level set scan statistic

1 Introduction

A common task of geospatial monitoring is to delineate zones that exhibit elevated risks over geographical regions. The problem is to identify concentrations of events in time, space or both. This need is found in many fields of application, such as epidemiology, landscape ecology, environmental analysis, image processing, economics, and public policy. This is essentially the problem of clustering with the added stipulation or constraint that the components of a cluster must be spatially connected in the sense of adjacency. We discuss two methods of imposing this spatial constraint on clusters. The first method (I) does retrospective screening on the results of a clustering algorithm to identify connected components. The second method (II) applies the spatial constraint as the clustering algorithm progresses. The first method has the advantage of separable programming for the spatial screening without requiring modification of existing software for clustering, and of using the same spatial screening software with different clustering algorithms. However, the two methods often yield different results in practice. We demonstrate this inconsistency for two popular clustering techniques, single-linkage and K-means, then consider the upper level set (ULS) scan (Patil and Taillie 2004) as a special clustering approach that produces the same results using both methods.

2 Background

2.1 Clustering with constraints

Constrained clustering is defined by Legendre (1987) as a part of the family of clustering methods whose purpose is to delimit homogeneous regions on a univariate or multivariate surface, by forming blocks of pieces that are also adjacent in space or in time (Legendre 1987). The most common applications are spatial, temporal or both spatial and temporal constrained clustering. A variety of studies have been done regarding the methods and applications of clustering with constraints. Legendre (1987) has reviewed the development of constrained clustering among different disciplines, including economics, soil science, ecology, and so on. He has compared several clustering methods, from single to complete linkage, and concluded that the clusters derived using different linkage clustering methods are quite similar to one another when the constraint is applied (Legendre 1987). The difference of constrained clustering and unconstrained clustering is also explained by Legendre (1998). A number of programs have been written, such as the program of Legendre (1987), Lebart (1978), and De Soete (1987).

This paper focuses on spatially constrained clustering. Specifically, the clustering is spatially constrained if the cells comprising a cluster are adjacent to each other in addition to having similarity with regard to the measured variables.

2.2 Spatially constrained clustering

Spatially constrained clustering is useful in applications in many fields, such as landscape ecology, geospatial environmental statistics, image processing, geospatial economics, public policy in geospace, geography of disease, and almost every field of study where spatial continuity is relevant. In ecology, for example, if one wishes to discover clusters of species richness, a constraint of spatial adjacency needs to be added, since the contagion of biological growth cannot be ignored. In health studies, if one is interested in finding clusters of a certain disease (i.e. lung cancer), it is desirable to take into consideration geographical adjacency, since the disease may be caused by an environmental pollutant, which is often spatially autocorrelated. It is essential in finding out the pattern of how the disease spreads out.

In general, spatially constrained clustering is accomplished using two methods (Urban 2004):

I. Spatial constraint is applied after conventional clustering based on similarity.
II. Spatial constraint is applied as clustering progresses.

In constraining method I the analysis is conducted in two stages, as described and illustrated by Legendre et al. (2002). At the first stage, the dataset is classified into different groups using regular clustering techniques; then spatial adjacency is examined at the second stage, where those adjacent members that are in the same group are combined into patches. On the other hand, constraining method II applies the spatial constraint progressively, asking the question whether or not two cells are adjacent for each of the iterations of the clustering algorithm.

Comparing the two methods, the result of constraining method I is more readily interpretable with regard to the variables, but it tends to give small patches and splinters. Constraining method II yields tighter clusters, but it is more problematic to interpret (Urban 2004). So, if one is more concerned about the similarity of cells, constraining method I is usually implemented. If one is more concerned about the connectivity of cells, then constraining method II is usually applied. The two methods often do not yield the same result. As far as programming is concerned, constraining method I has programs ready to use and it is more computationally efficient because it does not have to deal with the spatial constraint until the last step. Legendre, Vaudor, and Casgrain (2005) provide a number of programs in R that are useful for contiguity-constrained clustering. Programs are available for hierarchical agglomerative linkage clustering with spatial contiguity constraint and for K-means non-hierarchical constrained clustering. Usually, the two methods do not yield the same result.

An example serves to show how the two constraining methods can give different clusters for the same set of observations. The relationships of five observational units (A, B, C, D, E) are specified in Table 1, with dissimilarity as distance in Table 1a and spatial connectivity in Table 1b. All five observational units are spatially connected except for B and D.

Suppose that we use clustering by single-linkage, whereby the similarity of two cluster groupings is measured by the shortest distance between a member of one group and a member of the other group. Let the progress of clustering be agglomerative, whereby each data case is a singleton cluster at the outset. The algorithm proceeds as follows:

Table 1 Dissimilarity as distance and spatial connectivity for five illustrative data cases (A–D). A value of 1 in the connectivity matrix indicates spatial adjacency

	A	B	C	D	E
(a) Distance matrix					
A	0				
B	19	0			
C	13	17	0		
D	16	15	19	0	
E	21	20	12	18	0
(b) Connectivity matrix					
A	0				
B	1	0			
C	1	1	0		
D	1	0	1	0	
E	1	1	1	1	0

1. Start with N clusters with distance matrix D.
2. Search for nearest clusters x and y with the smallest distance d_{xy}.
3. Merge clusters x and y to form (xy) by
   - deleting rows and columns of x and y
   - adding cluster (xy)
   - compute new distance $\mathrm{d}_{(\mathrm{xy})\mathrm{z}} = \min(\mathrm{d}_{\mathrm{xz}}, \mathrm{d}_{\mathrm{yz}})$
4. Repeat steps 2–3, $N - 1$ times.

Applying constraining method I, the algorithm proceeds as follows:
Start with 5 clusters A, B, C, D and E. From the distance matrix we find that the points E and C are the closest and we merge them together. With E and C merged together the new distance matrix is as follows:

Distance matrix

	(EC)	A	B	D
(EC)	0			
A	13	0		
B	17	19	0	
D	18	16	15	0

Repeating these steps we find that we are able to form clusters (ACE) and (BD). The next step would merge them together to form the cluster (ABCDE). If we stop here and apply the spatial constraints to obtain the final clusters, we see from the connectivity matrix that objects A, C, and E form one cluster since they are connected. Objects B and D are not connected, so they remain separate. Hence, the final clusters using method I are (ACE), (B), and (D).

Now suppose we use constraining method II. During each step of clustering we inspect the connectivity matrix to determine whether the objects identified by the clustering algorithm should be joined. Method II also obtains the cluster (ACE) as a consequence of spatial connectivity. Consider the distance matrix at some intermediate stage of clustering.

Distance matrix

	(ACE)	B	D
(ACE)	0		
B	17	0	
D	16	15	0

The next step would be to merge B and D together, but since B and D are not spatially connected, we cannot merge them. Since now (ACE) and D are the spatially connected objects that have the least distance, they are merged. Hence, the clusters obtained using method II are (ACDE) and (B). This example shows that methods I and II can give different clusters.

In a K-means algorithm also, the two methods yield different results in most cases. Constraining method I will take the K clusters obtained by K-means clustering and apply the constraints to them. This may take one or more clusters that have disconnected objects and further subdivide them into even smaller connected subgroups. Hence, the numbers of groups produced by method I may be larger than K whereas method II produces K clusters as long as there are at least K data cases that differ in their observational values. For example, suppose we want to divide 4 objects A, B, C and D into $K = 2$ clusters such that objects within the first cluster are closer to each other than they are to the objects in the second cluster (2-means clustering). Suppose that objects A and B are not connected. Using the Euclidean distance, the 2-means clustering (method-2) obtains two clusters. Suppose the 2-means clustering method produces the clusters (AB) and (CD). Applying the connectivity constraint to these two clusters will produce A, B, and (CD) as the objects A and B are not connected. Hence constraining methods I and II may produce different numbers of clusters.

3 ULS scan

Upper level set (ULS) scan as discussed in Patil and Taillie (2004) can be considered as a special approach of spatially constrained clustering. ULS scan deals with a region $\boldsymbol{R}$ of two-dimensional space that is tessellated or subdivided into cellular areas (which will be denoted by $\boldsymbol{a}_i$). A non-negative value $\boldsymbol{G}_i$ is available as response (level) data for each cell, comprising a step surface. The ULS scan identifies spatially connected clusters of cells that have an elevated response relative to a progressively declining threshold level that scans from the maximum level to the minimum level. For a given level $\boldsymbol{g}$ of the scanning threshold, such a spatially connected cluster of elevated cells will be referred to as a **zone** denoted by $\boldsymbol{Z}_j$ and the set of all $\boldsymbol{Z}$ is denoted by $\boldsymbol{\Omega}$. Again for a specified level $\boldsymbol{g}$ of the scanning threshold, the set of all cells with response level equal to or exceeding $\boldsymbol{g}$ is the upper level set (ULS)

$$\boldsymbol{U}_g = \{\boldsymbol{a}_i : \boldsymbol{G}_i \geq \boldsymbol{g}\}.$$

Upper level sets for two particular threshold levels $\boldsymbol{g}$ and $\boldsymbol{g}'$ of a hypothetical surface are illustrated schematically in Fig. 1. The ULS for level $\boldsymbol{g}$ consists of zones $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ since these are the portions of the surface that equal or exceed the $\boldsymbol{g}$ threshold level. Likewise, the ULS for level $\boldsymbol{g}'$ consists of zones $\boldsymbol{Z}_4, \boldsymbol{Z}_5$ and $\boldsymbol{Z}_6$ since these are the portions of the surface that equal or exceed the $\boldsymbol{g}'$ threshold level.

Figure 1 also illustrates the three ways in which zones can change as the threshold level shifts. Zones $\boldsymbol{Z}_1$ and $\boldsymbol{Z}_2$ expand and merge to create zone $\boldsymbol{Z}_4$. Zone $\boldsymbol{Z}_3$ under-

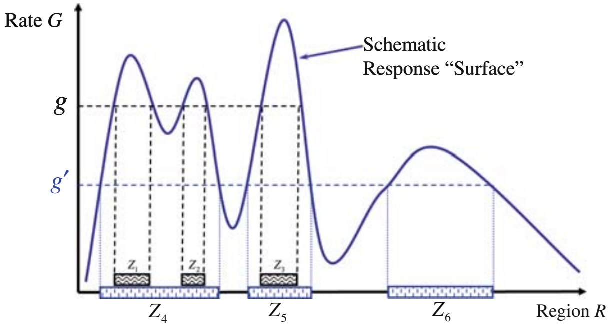


Fig. 1 Schematic response "surface" showing upper level sets (ULS) for two levels $\boldsymbol{g}$ and $\boldsymbol{g'}$ of scanning threshold. The ULS for threshold level $\boldsymbol{g}$ consists of zones $\boldsymbol{Z}_1, \boldsymbol{Z}_2$ and $\boldsymbol{Z}_3$ as the portions of the surface with level at or exceeding the threshold. The ULS for threshold level $\boldsymbol{g'}$ consists of zones $\boldsymbol{Z}_4, \boldsymbol{Z}_5$ and $\boldsymbol{Z}_6$ as the portions of the surface with level at or exceeding that threshold

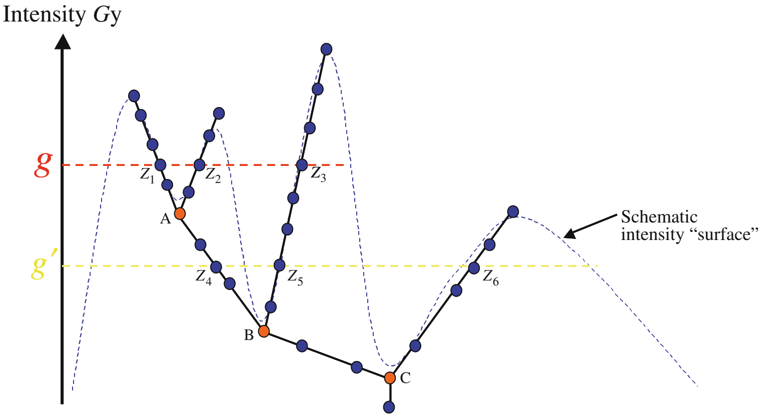


Fig. 2 Schematic of zonal ULS tree for example "surface" of Fig. 1. Small filled circles are nodes representing spatially connective zones at the respective scan level, with scan level $\boldsymbol{g}$ matching that in Fig. 1. Arcs (lines) linking nodes indicate formative relations among the zones at different scan levels, with a higher node linked to a lower node being spatially contained in the zone of the lower node. The dotted line represents the schematic "surface"

goes simple expansion to become zone $\boldsymbol{Z}_5$. Zone $\boldsymbol{Z}_6$ is newly emergent. It may aid visualization to think of the downward scanning threshold as a declining water level and the upper level sets as being emergent islands of land surface. Modarres and Patil (2006) provide an algorithmic explanation of ULS scan statistic along with more implementation details.

The collective action of the scanning threshold can be captured in a zonal ULS tree as shown schematically in Fig. 2. The nodes of this ULS tree represented by small filled circles in Fig. 2 are the spatially connective zones, positioned vertically according to level of scanning threshold. Zones have discrete vertical positions since there can be no more levels of occurrence than there are cells in the tessellation, and only those levels that actually occur in the data are relevant to the scan. The arcs (lines) linking the nodes indicate the formative relations among the zones at different scan levels, with a higher node linked to a lower node being spatially contained within the zone of the lower node.

The set of zonal nodes transected horizontally at a particular scan level is comprised of spatially connective (constrained) clusters defined by that level of the scan. There are three possibilities of formative relations between successive scan levels:

(1) Two or more connected components coalesce into one as a crotch of the ULS tree. This occurs when one or more cells for the lower scan level form a bridge between zones of the higher scan level.
(2) An existing connected component (zone) for the higher scan level grows in size for the lower scan level placing nodes along a branch of the ULS tree.
(3) A new connected component (zone) is formed at the lower scan level that did not exist at the higher scan level, thus being the terminal of a new branch on the ULS tree.

The zonal ULS tree constitutes a formal mathematical graph with members of $\boldsymbol{\Omega}_{\mathrm{ULS}}$ as nodes, and two nodes $Z, Z' \in \Omega_{\mathrm{ULS}}$ being joined by an edge if:

(i) Z is a *proper* subset of Z', written $Z \underset{\neq}{\subset} Z'$.
(ii) There is no node $W \in \Omega_{\mathrm{ULS}}$ such that $Z \underset{\neq}{\subset} W \underset{\neq}{\subset} Z'$.

For parallelism with usual conventions of clustering, we can define a distance matrix between cells for a specific threshold level. The similarity feature of interest is whether or not two cells are both in the upper level set or both in the lower level complement set. If two cells are both in the upper level set or both in the lower level set, then their distance is zero. Alternatively, if one cell is in the upper level set and the other in the lower level set, then their distance is one. Therefore, a distance matrix for a particular position in the scan cycle is a zero-one binary matrix. It should also be evident that lower levels of a tessellated step surface can be explored in equivalent manner.

4 Constraining method I and method II for ULS

We explore retrospective versus progressive constraints comparatively for ULS scan in terms of an example. The example uses the tessellated mapping of a two-dimensional geo-space shown in Fig. 3, with the numbers inside the polygons being just identifiers for the polygonal cells. A companion cellular ULS tree for the example is shown in Fig. 4. To focus attention further on the cellular ULS tree, we pursue the example without giving explicit values for the surface data.

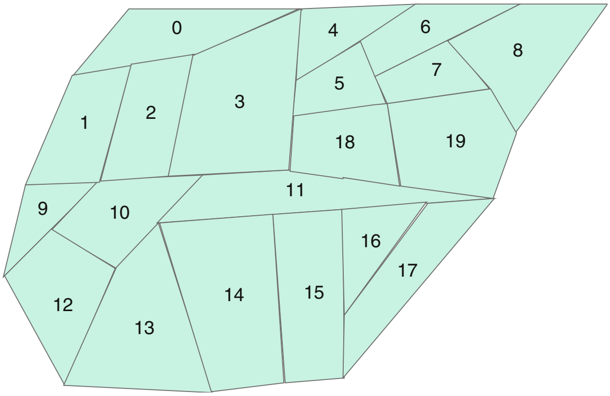
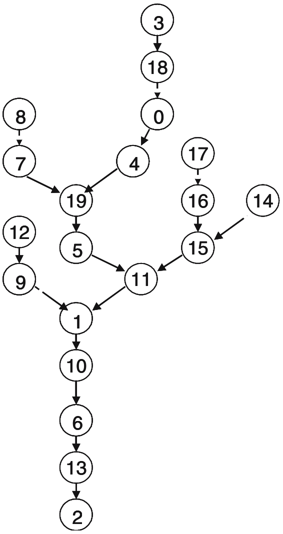


Fig. 3 Mapping of tessellated cells for example with cell identification numbers shown inside cell polygons

Fig. 4 Cellular (as opposed to zonal) ULS tree for example mapping in Fig. 3, with numbers in nodes being identifiers of cells that join into upper level set connectivity relations according to link arrows. Thus, each branch or stem of the tree represents a vertically developing zone in this way of structuring the tree. Height positions in tree reflect relative levels (step down) of surface values



For constraining method I the connectivity will not be checked until the last step. In the first phase we only develop the upper level sets without organizing them into zones. The connectivity matrix is applied at the end to obtain the zones, which also structures the linkages of the cellular ULS tree across the levels. Figure 5 shows the level information on the left, and the ultimate cellular ULS tree on the right. Note that the presentation of Fig. 5 only shows the level at which a cell first enters the upper set. The cell then remains in the upper set for lower scanning levels of the threshold.

For constraining method II the connectivity of cells is examined at each step (scanning level of the threshold) as shown in Fig. 6. The numbers in brackets of Fig. 6 are cell identifiers for members of connected components (zones). For example, both cells 8 and 0 enter the upper set at scanning level 3, but they cannot be merged together as

- Level 1: 3
- Level 2: 18
- Level 3: 8 0
- Level 4: 7 4 17
- Level 5: 19 16 14
- Level 6: 12 5 15
- Level 7: 9 11
- Level 8: 1
- Level 9: 10
- Level 10: 6
- Level 11: 13
- Level 12: 2

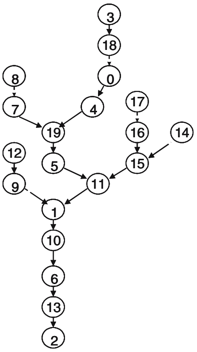


Fig. 5 Application of constraining method I to the example of Figs. 3 and 4, whereby the spatial constraint is imposed at the end. The scanning level at which cells enter the upper set is shown on the left, remaining in the upper set for lower scans. The cellular ULS tree resulting from retrospective spatial constraint is shown on the right

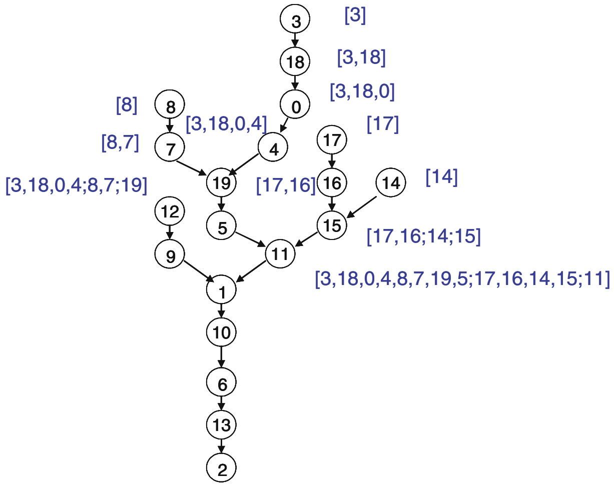


Fig. 6 Application of constraining method II to example of Figs. 3 and 4, whereby the spatial constraint is imposed progressively. The numbers in brackets are identifiers of cells comprising connected components (zones) at the respective levels

a connected component since they are not spatially adjacent. Thus, a new zone grows from cell 8 into the next lower scan level.

The two constraining methods give the same result for the example, and this holds in general for the ULS scan approach. It does not matter for ULS scan at what stage the spatial constraint is applied, either at the end or progressively, the result is the same.

5 ULS scan statistic

To this point in the consideration there has been no restriction on the type of data $\boldsymbol{G}$ for the surface level, and also no mention of test of significance for the zones of the upper level set. Following Kulldorff and Nagarwalla (1995), however, it is possible to consider significance of a spatial scan statistic when working with level data consisting of incidence rates. Such spatial scan statistics are based on null models of binomial or Poisson random processes with Monte Carlo generation of distributions for testing significance under the null hypothesis of random occurrence for upper level zones.

To motivate a null model, assume that data are in the form of a count $\boldsymbol{Y}_i$ (non-negative integer) on each cell. In addition, a “size” value $\boldsymbol{A}_i$ is associated with each cell. The cell sizes $\boldsymbol{A}_i$ are regarded as known and fixed, while cell counts $\boldsymbol{Y}_i$ are assumed to be independent random variables. The level variable is then the rate $\boldsymbol{G}_i = \boldsymbol{Y}_i/\boldsymbol{A}_i$.

Two distributional settings are commonly studied:

- **Binomial:** $\boldsymbol{A}_i = \boldsymbol{N}_i$ is a positive integer and $\boldsymbol{Y}_i \sim \text{Binomial}(\boldsymbol{N}_i, \boldsymbol{P}_i)$ where $\boldsymbol{p}_i$ is an unknown parameter attached to each cell with $0 < \boldsymbol{p}_i < 1$.
- **Poisson:** $\boldsymbol{A}_i$ is a positive real number and $\boldsymbol{Y}_i \sim \text{Poisson}(\boldsymbol{\lambda}_i, \boldsymbol{A}_i)$ where $\boldsymbol{\lambda}_i > 0$ is an unknown parameter attached to each cell.

Each distributional model has a simple epidemiological interpretation. For the binomial, $\boldsymbol{N}_i$ people reside in cell $\boldsymbol{a}_i$ and each can contract a certain disease independently with probability $\boldsymbol{p}_i$. The cell count $\boldsymbol{Y}_i$ is the number of afflicted people in the cell. For the Poisson, $\boldsymbol{A}_i$ is the size (perhaps area or some adjusted population size) of the cell $\boldsymbol{a}_i$, and $\boldsymbol{Y}_i$ is a realization of a Poisson process of intensity $\boldsymbol{\lambda}_i$ across the cell. In each scenario, it is assumed that spatial variability can be accounted for by cell-to-cell variation in the model parameters. The spatial scan statistic seeks to identify "hotspots" of cells that have elevated response compared with the rest of the region. Elevated response means large values for the rates as opposed to raw counts $\boldsymbol{Y}_i$. Cell counts are thus adjusted for cell sizes before comparing cell responses. The scan statistic also accommodates other adjustments, such as for age or for gender. A candidate zone to be tested should not be excessively large—for, otherwise, the zone instead of its exterior would constitute background. This latter restriction is generally achieved by limiting the search for hotspots to zones that do not comprise more than, say, 50% of the region. The number of zones is small enough for practical maximum likelihood search.

The ULS method has been applied to detect clusters of West Nile Virus in the state of Ohio. West Nile Virus is transmitted by mosquitoes and birds. The 3D map of disease rate is shown in Fig. 7. Five shades are distributed in quintiles. The top two clusters are detected using the ULS scan. Each cluster consists of counties that are geographically connected. This result is expected since the clusters are restricted to be composed of only spatially adjacent counties. The candidates that are statistically significant are claimed as real hotspots.

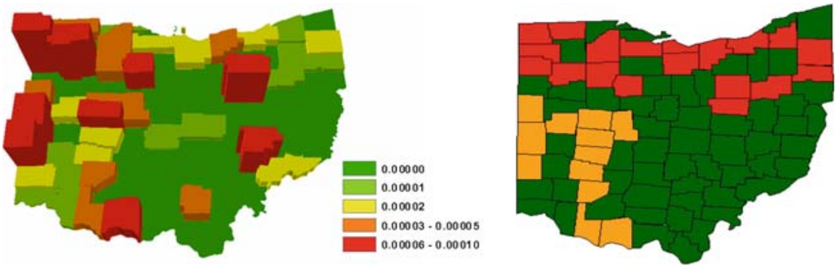


Fig. 7 Ohio case study of ULS scan statistic for West Nile virus. A 3D representation of incidence rates is shown on the left and two significant hotspots on the right

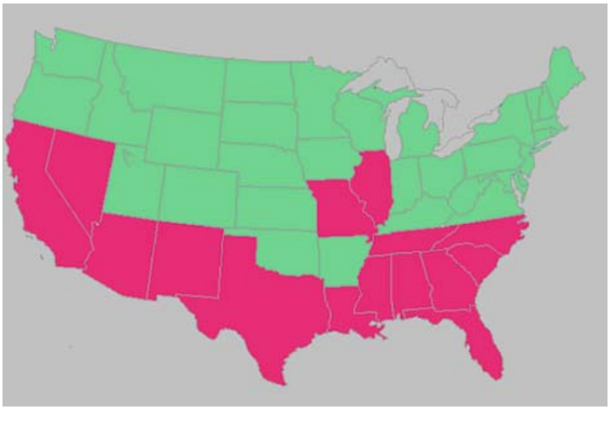


Fig. 8 Elevated zone of robbery in southern states extending into Missouri and Illinois

6 ULS scan on multivariate data by spatial coincidence

So far we have seen how ULS works on univariate data. But what happens when the data under consideration are multivariate. In such a scenario ULS is operated as many times as the response dimensions of the data, with every individual run of ULS operating only on one response variate. A multivariate result for upper level zone(s) is then obtained as the spatial intersection of the clusters for individual variates. For example if we have data with every cell in the tessellation having two response variables, then we run ULS twice. The first run will take into account the first response variable for every cell and find the spatially connected clusters. The second run will take into account the second response variable and find the spatially connected clusters. Finally the clusters are those that are the intersection of the clusters obtained by the two runs of ULS. As in the univariate case, constraining method I and method II will provide the same result.

As a working example consider crime data for US states. Every state has two observations, the count of robbery in that state and the count of murders committed in that state. The aim is to cluster those states that have a high incidence of both

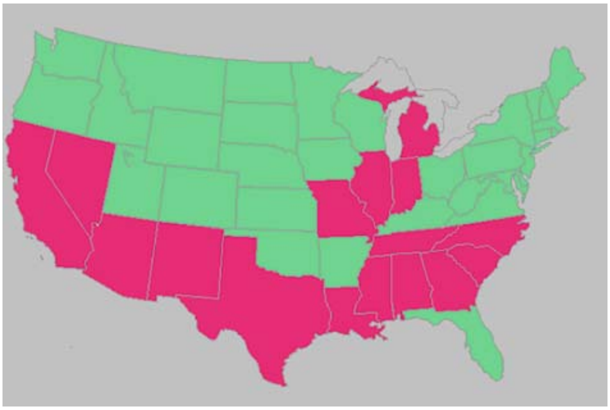


Fig. 9 Elevated zone of murder in south and Missouri, Illinois, Indiana and Michigan

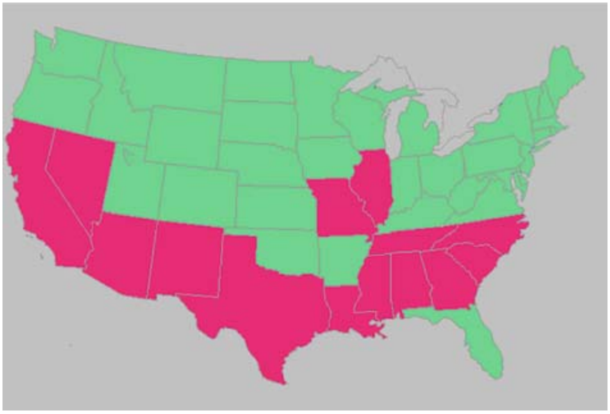


Fig. 10 Jointly elevated zone of robbery and murder, intersecting zones of Figs. 8 and 9

robbery and murder. We run ULS twice, once taking into account the response variable for robbery and again taking into account the response variable for murder. For simplicity we show only one elevated cluster in each case. A first run of ULS takes into account the response variable for robbery (Fig. 8), with an elevated zone across southern states and extending up into part of the Midwest. A second run of ULS takes the response variable for murder into account (Fig. 9), with a similar elevated zone but also extending on up into Indiana and Michigan. Then the spatially connected zone having high incidence of both robbery and murder consists of the states that are identified as an intersection of the two runs of ULS (Fig. 10).

7 Conclusion

We are interested in identifying clusters of events in time, space or both. In many cases, usual clustering analysis is not sufficient and one must account for the spatial structure of the data. For example, it is necessary to bring in spatial constraints in the presence of spatial autocorrelation in the process generating the response, so that geographical continuity is taken into consideration. With the constraint of adjacency, clusters are formed only if the area at risk is locally homogeneous.

We discuss two methods of obtaining spatially constrained clusters. Constraining method I applies the constraint after the clustering algorithm obtains the clusters while method II applies the constraint as the clustering algorithm progresses. The two methods often yield different results under conventional clustering algorithms. We show that two popular clustering techniques, single-linkage and K-means, can yield different spatially constrained clusters under these methods.

We proceed to present an Upper Level Scan (ULS) method of clustering and demonstrate that it is a special case of constrained clustering. We explain how the ULS scan approach defines similarity as exceedance at each scan level and that the similarity matrix is binary. We also explain that connectivity is defined in terms of an adjacency matrix. The underlying tree structures of the ULS scan provide a levelwise graphical representation of the clusters akin to dendrograms. We show through several examples how the tree structure of ULS scan is formed under spatially constrained clustering and that ULS scan statistic yields the same spatial clusters under these two methods. We illustrate application of ULS scan statistics for incidence data. To use with multiresponse data, we discuss an approach where univariate ULS scan statistic is applied on the univariate responses and define the multivariate clusters as those in the intersection of univariate clusters. This and other methods of extending the univariate ULS to multivariate data are currently under investigation.

References

De Soete G (1987) Least squares algorithms for constructing constrained ultrametric and additive tree representations of symmetric proximity data. J Classifica 4:155–173

BoundarySeer (2005) Software for geographic boundary analysis. TerraSeer Ann Arbor MI

Everitt B, Landau S, Leese M (2001) Cluster analysis. Arnold, London, pp 161–164

Kulldorff M, Nagarwalla N (1995) Spatial disease clusters: detection and inference. Statistics in Medicine, 14:799–810

Lebart L (1978) Programme d'agrégation avec constraints (C.A.H. contiguïté). Cah Anal Donnée 3:275–287

Legendre P (1987) Constrained clustering. In: Legendre L (eds) Developments in numerical ecology, Springer-Verlag, Berlin. pp 289–307

Legendre P, Legendre L (1998) Numerical ecology. Elsevier, NY, pp 853

Legendre P, Ellingsen K, Bjornbom E, Casgrain P (2002) Acoustic seabed classification: improved statistical method. Can J Fish Aqua Sci 59(7):1085–1089

Legendre P, Vaudor A, Casgrain P (2005) The R package for spatial and multidimensional analysis. Software available at http://www.bio.umontreal.ca/Casgrain/en/labo/index.html

Modarres R, Patil GP (2006) Hotspot Detection with Bivariate Data. To appear in the J Stat Plan Infer

Patil GP, Taillie C (2004) Upper level set scan statistic for detecting arbitrarily shaped hotspots. Environ Ecol Stat 11:183–197

Urban DL (2004) Multivariate Analysis. Nonhierarchical Agglomeration. Spatially Constrained Classification. http://www.env.duke.edu/landscape/classes/env358/mv_pooling.pdf

Biographical Sketches

Dr. G. P. Patil is Distinguished Professor of Mathematical Statistics and Director, of the Center for Statistical Ecology and Environmental Statistics in the Department of Statistics at the Pennsylvania State University, University Park. PA. USA. He is Editor-in-Chief, of *Environmental and Ecological Statistics*. He is Principal Investigator of a five-year NSF Digital Government Program Project on Surveillance GeoInformatics for Hotspot Detection, Prioritization, Early Warning, and Sustainable Management.

R. Modarres is an Associate Professor of Statistics in the Department of Statistics at The George Washington University.

Wayne L. Myers is Professor of Forest Biometrics in the School of Forest Resources and Director of the Office for Remote Sensing and Spatial Information Resources (ORSSIR) in the Penn State Institutes of Environment (PSIE) at the Pennsylvania State University. His research interests include remote sensing, geographic information systems, landscape ecology, biodiversity, and forest resources inventory.

Pushkar Patankar is a graduate student in the Department of Computer Science and Engineering at The Pennsylvania State University at University Park, PA.