



Center for **S**tatistical **E**cology and **E**nvironmental **S**tatistics

Hotspot Detection with Bivariate Data

By Reza Modarres and G.P. Patil

¹Department of Statistics, George Washington University
Washington, DC 20052, USA

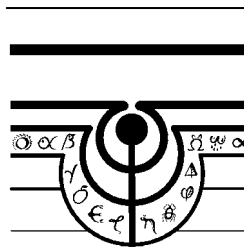
²Department of Statistics, The Pennsylvania State University
University Park, PA 16802, USA

This material is based upon work supported by (1) the National Science Foundation under Grant No. 0307010, (ii) The United States Environmental Protection Agency under Grant No. RD-8324401.

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.

[Invited Paper for the Journal of Statistical Planning and Inference, Centennial Volume, S.N. Roy]

Technical Report Number 2006-0501
TECHNICAL REPORTS AND REPRINTS SERIES
May 2006



Department of Statistics
The Pennsylvania State University
University Park, PA 16802

G. P. Patil
Distinguished Professor and Director
Tel: (814)865-9442 Fax: (814)865-1278
Email: gpp@stat.psu.edu

<http://www.stat.psu.edu/~gpp>
<http://www.stat.psu.edu/hotspots>

[DGOnline News](#)
[Environmental and Ecological Statistics-Springer](#)

Hotspot Detection with Bivariate Data

REZA MODARRES¹ and G. P. PATIL²

¹ *Department of Statistics, The George Washington University
Washington DC 20052, USA*

² *Department of Statistics, The Pennsylvania State University
University Park, PA, 16802, USA*

Abstract

The Upper Level Scan Statistic (ULS), its theory, design and implementation, and its extension to the bivariate data are discussed. We provide the ULS-Hotspot algorithm that obtains the response rates, maintains a list of connected components at each level of the rate function and yields the ULS tree. The tree is grown in the immediate successor list, which provides a computationally efficient method for likelihood evaluation, visualization and storage. An example shows how the zones are formed and the likelihood function is developed for each candidate zone. The general theory of bivariate hotspot detection is explained, including the bivariate binomial model, the multivariate exceedance approach, and the bivariate Poisson distribution. We propose the Intersection method that is simple to implement, using a univariate hotspot detection methods. We study the sensitivity of the joint hotspots to the degree of association between the variables. An application for the mapping of crime hotspots in the counties of the state of Ohio is presented.

MSC: primary 62H10; 62H15; 62F03; secondary 62P12

Key Words: Hotspots, Geoinformatic Surveillance, Spatial Scan Statistic, Upper Level Set Scan Statistic, Bivariate, Crime Mapping.

1 Introduction

Geoinformatic surveillance for detection of spatial and temporal hotspots is a declared need for the modern society. A hotspot refers to a cluster of events in space and time with elevated responses, an unusual occurrence and an oddity, such as an outbreak, or any departure from

¹This work was completed while the author was on sabbatical at the Center for Statistical Ecology and Environmental Statistics of the Pennsylvania State University. E-Mail: Reza@gwu.edu.

²This material is based upon work supported by the National Science Foundation under Grant No. 0307010 and the US EPA under Grant No. RD-8324401. 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. E-Mail: gpp@stat.psu.edu.

a geo-referenced set of prior expected responses. The causes are varied and maybe willful, natural or accidental. The need concerns development of statistical methods for detection of hotspots and software infrastructure. What is particularly needed is fast detection of arbitrarily shaped hotspots. Identification of critical spots (coldspots have depressed rates and are treated similarly), evaluation of the significance of the found cluster and assessment of covariates form the skeleton of a hotspot detection method and the associated software. Several techniques for the detection of hotspots have appeared in the literature, including the spatial scan statistic, SaTScan, (Kulldorff and Nagarwalla, 1995; Kulldorff, 1997) and the upper level set, ULS, scan statistic (Patil and Taillie, 2004). SaTScan is available on the Internet and the ULS scan is under development (Patil, Modarres, and Patakar, 2005).

It is our aim in this article to discuss the ULS scan statistic, its theory, design and implementation, and its extension to the bivariate case. We will discuss the theory of scan statistic in section 2. Implementation issues are discussed in section 3. We extend the method to the bivariate case in section 4 where we study the bivariate binomial and bivariate Poisson models and other test statistics. We describe a method of conducting a sensitivity analysis for the hotspots in section 5. In an application, we describe the detection of crime hotspots. The last section is devoted to summary and concluding remarks.

2 Theory of ULS Scan Statistic

The ULS scan statistic is composed of three main components. First, there is geometry of the scanning area to consider. The scanning region R of the Euclidian space is partitioned into cells. For example, a region maybe subdivided by counties, by postal zip codes, or other methods of forming boundaries. Each subdivision is commonly referred to as a cell. The observed responses in each cell and their sampling distribution under the null hypothesis form the second component of the ULS. At each cell, a , we have available a non-negative count Y_a and a known size A_a . Two commonly studied models, the binomial and the Poisson, are used to model the cell counts (the responses). Under a binomial model, the fixed size A_a represents the number N_a of organisms (individuals, animals, plants, pixels, etc) in cell a , each with a certain attribute (disease) independently with probability P_a . The cell count Y_a represents the number of individuals with that attribute. Hence, $Y_a \sim \text{Binomial}(N_a, P_a)$. In a Poisson model, A_a often represents the area or the population size of cell a and Y_a is a Poisson process with intensity λ_a across the cell.

Both models assume that the responses are independent and that the spatial variability is explained by the cell-to-cell variation of the model parameters. One can also model continuous responses by modeling their means and variances. Patil and Taillie (2004) consider gamma and lognormal distributions. Further research on the use of other continuous distributions such as beta, Pareto or Weibull is needed. The first two components of ULS are

shared by other scan statistics such as SaTScan. The third component concerns the shape and size of the scanning window and directly relates to the efficiency and the power of the scanning algorithm. Unlike SaTScan, which uses a circular or an elliptical scanning window, the scanning window in ULS is data-driven and is determined by the piece-wise constant surface of the responses over the region. This surface will allow for prudent selection of connected cells (candidate zones) from the tessellation. A zone Z is a set of connected cells in the region R . The set of all zones is denoted with Ω (see Figure 1).

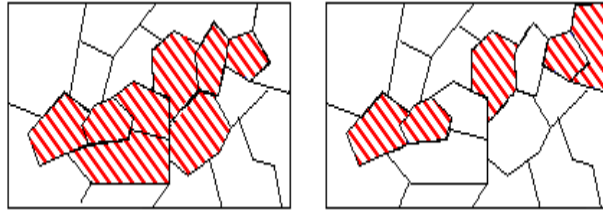


Figure 1: A tessellated region. The collection of shaded cells on the left is connected, hence a zone Z in Ω . The collection on the right is not connected.

The ULS scan statistic searches for clusters of cells or candidate zones that exhibit elevated responses relative to the rest of the region. The rates $G_a = Y_a/A_a$ are used by ULS scan statistic to form candidate zones. In practice, the search for hotspots is limited to zones that are not very large and at most comprise fifty percent of the population. A hotspot is a zone Z whose likelihood of occurrence relative to the likelihood of the expected responses is too small to attribute to chance variations. For example, under a binomial model, one may state the null and the alternative hypotheses as $H_0 : P_a$ is constant across all cells in R (no hotspots) and $H_1 : \text{there is a zone } Z \text{ such that for } P_1 > P_0$,

$$P_a = \begin{cases} P_1, & \text{if } a \in Z, \\ P_0 & \text{if } a \in Z' = R - Z. \end{cases} \quad (1)$$

Under the alternative hypothesis, the zone Z is an unknown model parameter while under the full model $H_0 \cup H_1$, the unknown parameters are the zone $Z \in \Omega$, P_0 , and P_1 . For a given set of connected cells, candidate zone Z , the profile likelihood over the space $0 < P_0, P_1 < 1$ is $L(Z) = \max L(Z, P_0, P_1) = L(Z, \hat{P}_0, \hat{P}_1)$. Even though there are a finite number of zones in Ω , the number is often computationally formidable. The connectivity requirement reduces the size of the search space, however, maintaining connectivity when forming the candidate zone Z can be an added burden. In practice, an exhaustive search over the connectivity restricted set of zones may still not be possible and a further reduction of the search space is required. The innovative aspect of the ULS scan statistic is the way the observed responses are used to order the search space. The ULS scan statistic restricts the search of zones Z with elevated rates to those $Z \in \Omega_{ULS}$ where Ω_{ULS} consists of all connected components of

all upper level sets of the rate surface. Note that any parameter-space reduction methods will run the risk of excluding the MLE from the search space. In practice the full parameter space is so large that a sub-optimal solution may be acceptable as proposed by both SaTScan with reduced search space $\Omega_{SatScan}$, (Kulldorff and Nagarwalla, 1995), and Ω_{ULS} (Patil and Taillie, 2004). Both methods work well when the reduced parameter space contains the MLE or at least a good approximation to the MLE. This shortcoming of parameter-space reduction techniques is a trade-off with the excessive computer time a full search may require. The reduced parameter space Ω_{ULS} provides for a faster and adaptive detection method and distinguishes ULS from other hotspot detection techniques. The resulting method is fast because the cardinality of Ω_{ULS} is at most the number of cells in R . Furthermore, Ω_{ULS} has the structure of a tree, which is useful for visualizing the hotspot clusters that form on the tree (see Figure 5 and the next section).

The rates G_a define a piece-wise constant surface over the region. The reduced search space Ω_{ULS} is the set of all connected components of all the upper level sets (ULS) of this surface. The rates specify a data-driven function that associates the cell a to its rate. This is a step-function with a finite number of levels and each level g determines a ULS $U_g = \{a : G_a \geq g\}$. There is an underlying tree structure associated with Ω_{ULS} . The nodes of the tree are zones in Ω_{ULS} and the resulting ULS-tree is defined on the hills and valleys of the rate function as zones of connected cells are formed. The leaf nodes correspond to a single cell where the response rate is typically the local maximum. Figure 2 shows the response surface at two response levels and how ULS is determined. Figure 3 shows the associated ULS-tree. The presented surface in Figure 2 and 3 are schematic surfaces. Figure 3 provides a picture of the ULS-tree formed on the hills and valleys of the rate function. The observed surface is a city skyline. We note, in section 3, that the response rates are ordered and the cell with the maximum response rate will initialize a candidate zone. The order is well-defined when the response rates are distinct. The ties can be broken randomly or in the order that the cells are scanned.

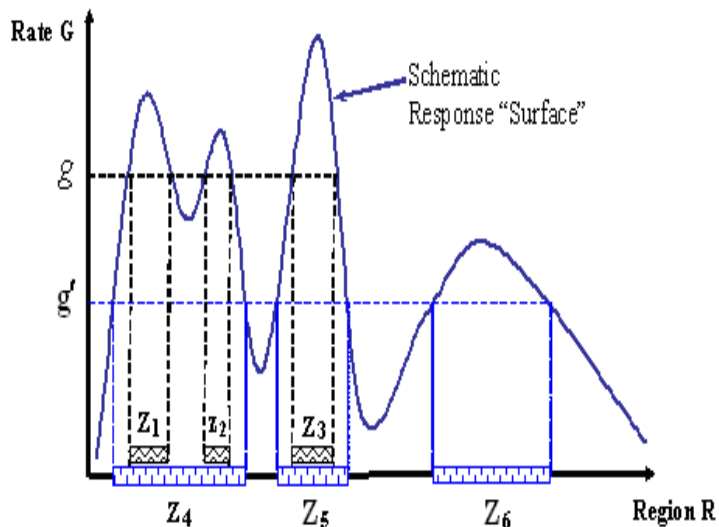


Figure 2: Schematic response surface with two response levels, g and g' . The upper level set determined by g has three connected components, Z_1 , Z_2 , and Z_3 ; that determined by g' has Z_4 , Z_5 , and Z_6 as its connected components. The diagram also illustrates the three ways in which connectivity can change as the level drops from g to g' : (i) zones Z_1 and Z_2 grow in size and eventually coalesce into a single zone Z_4 , (ii) zone Z_3 simply grows to Z_5 , and (iii) zone Z_6 is newly emergent.

The zones in Ω_{ULS} are candidate hotspots since they are portions of upper level sets of the response rate. A tree structure can be defined on the reduced parameter space Ω_{ULS} . The nodes of the tree are the members of Ω_{ULS} , i.e., the candidate zones. Two nodes $Z, Z' \in \Omega_{ULS}$ are joined by an edge if

- (i) Z is a proper subset of Z' , written as $Z \subsetneq Z'$.
- (ii) There is no node $W \in \Omega_{ULS}$ such that $Z \subsetneq W \subsetneq Z'$.

This tree is called the ULS-tree; its nodes are the zones $Z \in \Omega_{ULS}$ and are therefore collections of vertices from the abstract graph. The root node consists of all vertices in the abstract graph. If one thinks of the tessellated surface as a landform, then initially the entire surface is under water. As the water level (the response rate) drops, more of the landform is exposed. At each water level, the cells are either exposed or unexposed. The newly exposed cell is either a local maximum (a new island emerges), or joins an existing island that increases in size, or a cell that joins two or more existing islands.

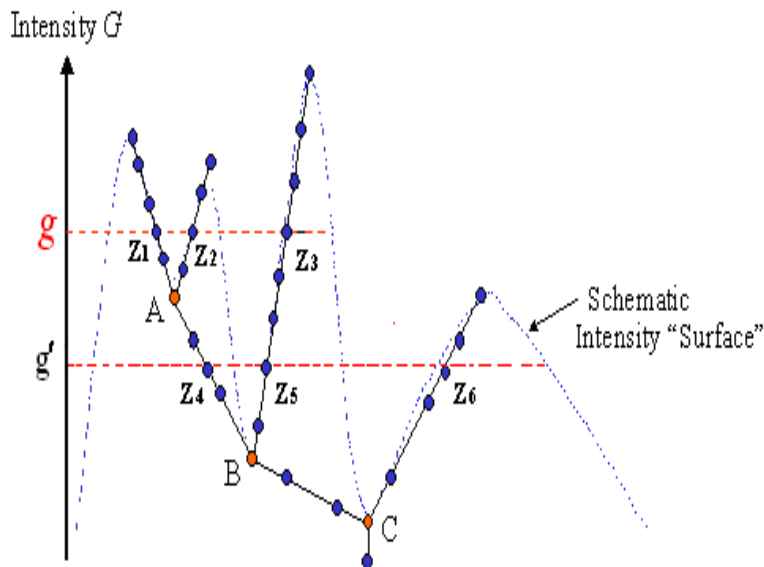


Figure 3. ULS connectivity tree for the schematic surface displayed in Figure 2. The four leaf nodes correspond to surface peaks. The root node represents the entire region. Junction nodes (A, B and C) occur when two (or more) connected components coalesce into a single connected component.

3 Design and Implementation Issues

The main components of the ULS scan statistic must be addressed from a design and implementation view. First, we need to represent the issue of the adjacency of the cells in R . ULS scan statistic uses an adjacency list to represent the adjacency structure of R . The adjacency relationships are provided by the user and consist of a listing of a cell label for each cell in the tessellation, followed by a list of its adjacent cells. A linked list implementation of the adjacency relationship will ease the merging and formation of new lists. The concept of adjacency is applied quite generally and includes cells that share a boundary, a point in common, or any other form of connection in which a cell receives input and is affected by another cell. For example, consider the flow from one cell to another in a drainage network. A consistency check on the adjacency list is performed to ensure its correctness by checking if cell a is adjacent to cell b , then cell b is adjacent to cell a .

Second, one needs to form the search space Ω_{ULS} and obtain the likelihood $L(Z, \hat{P}_0, \hat{P}_1)$ for each $Z \in \Omega_{ULS}$. To that end, one needs to keep track of the connectivity of the zone being formed at each level g of the ULS-tree. To obtain the ULS hotspots corresponding to the observed responses, the the following algorithm is implemented where cells are identified

using their cell labels. Cells are labeled consecutively with integers 0 through $K - 1$, where K is the numbers of cells in R . Cells with observed response rates are called exposed. Exposed is a list of logical values that identifies a cell as either exposed or unexposed depending on whether its rate is above or below a rate g . For example, in Figure 2, all cells in zones Z_1 , Z_2 and Z_3 are exposed at level g and all nodes below that rate are unexposed. The immediate successor of a cell a will point to the cell immediately below cell a in the ULS-tree. For example, in Figure 3, junction nodes A and B are immediate successors of nodes Z_1 , Z_2 and Z_4 , Z_5 , respectively. A terminal node refers to the last node on the list of connected components, $CCList$.

Algorithm: ULS-Hotspots

- Set up the adjacency list and input the rates as a function of the cell labels. If a is a node in the ULS-tree, then $cell = V[a]$ is the cell corresponding to that node. Similarly, $Node[a]$ is the rank of cell a with respect to its response rate. Note that $Node$ and V are inverse functions; thus, $Node[V[a]] = a$ for all a in R .
- Set $Node[V[a]] = a$ and $Exposed[a]$ to False for all cells a in R . The list element $S[a]$ will hold the immediate successor of cell a in the ULS tree. The list will be used to produce the ULS-Tree. Initially, $S[a] = a$.
- Place the rates in decreasing order. The list $V[a]$ will contain the corresponding cell labels. Set $CCList[a]$ to NIL for all cells a in R .
- For each cell a in R do the following:
 - Set $cell = V[a]$ and maintain a list of connected components $CCList$ for this cell. At each level g of the rate function, there will exist a $CCList$ of the connected components, $Connected[cell, g] = CCList[g]$. Since cell a is exposed at this time $Exposed[a]$ is set to True.
 - Scan the list of the cells b that are adjacent to $cell$. If the $CCList[b]$ is already exposed, then the list headed by a is merged with the list headed by b by keeping a at the top of the list. Otherwise, the algorithm will move to the next b .
 - Let U be the last cell of the list headed by b . Note that initially all cells are assumed unexposed and at the response rate drops, cells become exposed and form connected components.
 - Update the ULS tree: $S[Node[U]] = a$
 - For each cell x seen so far that is a terminal node, use $CCList[x]$ to calculate and store the likelihood of x .

- Order the likelihoods and identify the top hotspot as the node with the largest likelihood. Other hotspots can also be obtained in decreasing order of the list of the likelihood values provided that they do not have any cells in common with the previous hotspots.

Example: To illustrate the algorithm and obtain the ULS zones, consider the adjacency structure of the region shown in Figure 4 where the zones are depicted along with their likelihoods. Note that region is partitioned into 20 cells with cell labels 0 through 19. The cell labels of the first seven largest rates are (17, 3, 18, 8, 12, 14, 9). Hence, cell 17 is exposed first and becomes a zone by itself. Next, cells 3 and 18 form a zone whose likelihood is greater than that of cell labeled 17. Figure 5, constructed from the list S , shows how the ULS-tree is constructed at each level.

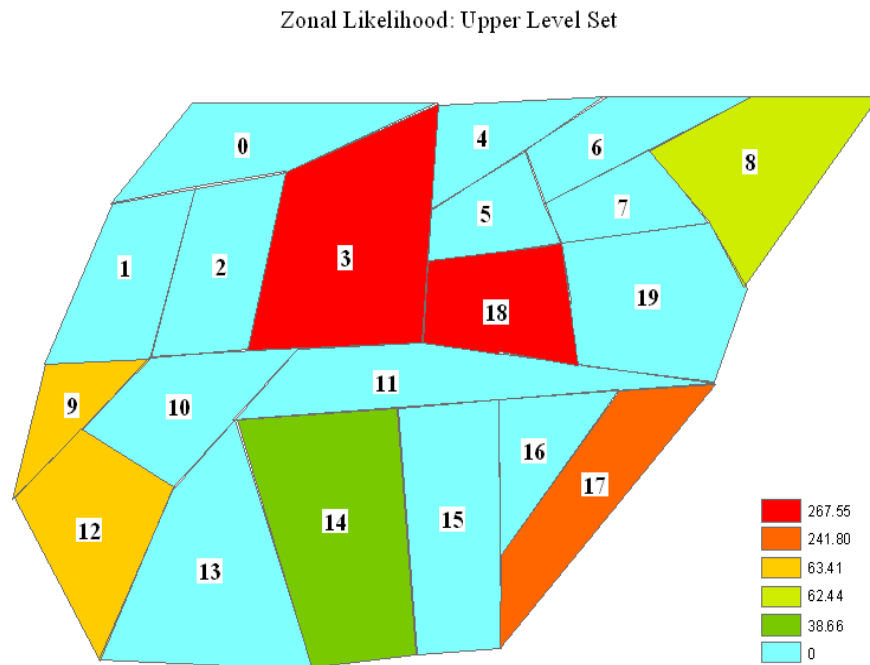


Figure 4: The adjacency structure of the region along with the likelihood of the zones.

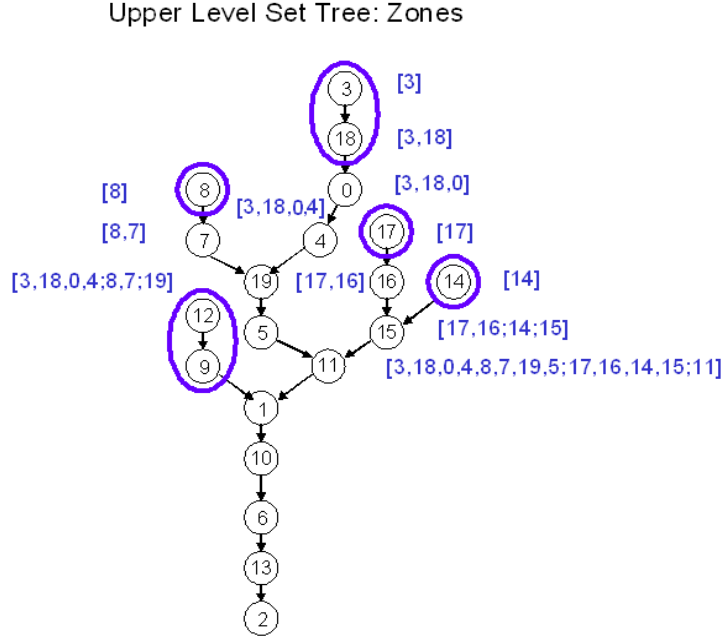


Figure 5: The Upper Level Set tree for adjacency structure in Figure 4 and constructed from the S list.

Note that the list S allows us to traverse the ULS tree starting from any given node downward to a terminal (root) node. The set of cells that are connected at level g of the tree is maintained in $Connected[cell, g]$. This matrix is maintained in ULS-Hotspots algorithm and is subsequently used to obtain the likelihood of the candidate zones. A linked list representation of the matrix will ease the memory burden caused by the simplicity of method. As part of the second component of ULS one must consider the form of the likelihood function and determine the significance of the top hotspots. Under a null hypothesis of no critical hotspot, the binomial likelihood is given by $L(null) = [\prod_a \binom{N_a}{Y_a}] P_n^{y_T} (1 - P_n)^{n_T - y_T}$ where $P_n = y_T/n_T$, the estimated disease rate under the null hypothesis, $y_T = \sum_a Y_a$, the total number of disease cases and $n_T = \sum_a N_a$, the total population size. Under the hypothesis that Z is a hotspot, the likelihood is $L(Z) = [\prod_a \binom{N_a}{Y_a}] P_Z^{y_Z} (1 - P_Z)^{n_Z - y_Z} P_{Z'}^{y_{Z'}} (1 - P_{Z'})^{n_{Z'} - y_{Z'}}$ where $P_Z = y_Z/n_Z$ is the estimated disease rate in zone Z , $P_{Z'} = y_{Z'}/n_{Z'}$ is the estimated disease rate obtained in $Z' = R - Z$, the population size outside zone Z is $n_{Z'} = \sum_{a \notin Z} N_a = n_T - n_Z$, the population size of zone Z is $n_Z = \sum_{a \in Z} N_a$ and the number of disease cases outside zone Z is $y_{Z'} = \sum_{a \notin Z} Y_a = y_T - y_Z$. The likelihood ratio is thus

$$LR = \frac{P_n^{y_T} (1 - P_n)^{n_T - y_T}}{P_Z^{y_Z} (1 - P_Z)^{n_Z - y_Z} P_{Z'}^{y_{Z'}} (1 - P_{Z'})^{n_{Z'} - y_{Z'}}}.$$

Calculate $LLR = -2\log(LR)$ for the zone with the largest likelihood, the maximum likeli-

hood estimate. To obtain the statistical significance of the top zone, it must be compared with the sampling distribution of LLR under the null hypothesis. The test statistic does not meet the regularity conditions necessary for it to have a chi-square distribution. To obtain its sampling distribution, one should use Monte Carlo simulation and condition on y_T with N_a and n_T known and fixed. The simulations are performed on independent and identical realizations of the data set given y_T . Conditioned on the sufficient statistics to eliminate the unknown parameters from the model under the null hypothesis, one obtains the hypergeometric under a Binomial model and the multinomial distribution under a Poisson model. Note that $L(null)$ does not depend on the zone Z and is calculated only once whereas $L(Z)$ must be recalculated for every new Z that is formed in the simulation. The number of simulated data sets to generate and analyze in order to estimate the null distribution and assess significance is typically in the range 999 to 9999. (Dwass, 1957).

Note that both univariate and bivariate models we discuss in this article have no spatial dependence. One can use one of several existing tests in the literature such as Morans I statistic (Moran, 1950) to examine if autocorrelations exist in the data. Here, we assume that spatial variability is accounted for by cell-to-cell variations in the model parameters. Ignoring autocorrelation tends to under estimation of variability and over estimation of the significance of the null hypothesis of no hotspots, leading to too many rejections. One can use a spatial autoregressive (SAR) model. Li, Calder and Cressie (2005) discuss testing for spatial dependence based on the SAR model.

4 Bivariate ULS Scan Statistic

The circle-based SaTScan and data-driven ULS scan statistic are designed to identify hotspots based on the elevated responses of one variable over the scan region. These techniques are appropriate for detecting univariate hotspots. What can be done when the data under consideration provides many correlated responses in each cell? The recent work on multivariate hotspot detection includes Burkom (2003) who discusses the use of multiple disparate data sources in biosurveillance and Wong, Moore, Cooper and Wagner (2003) who present an algorithm to detect disease outbreaks by searching for anomalous patterns in emergency department cases. Duczmal, Kulldorff and Mostashari (2005) use multiple data streams in a syndromic surveillance system to improve the detection power of a space-time permutation test statistic. Many techniques assume that the variables are independent so that the log-likelihood is a sum of the individual log-likelihoods or examine a scalar valued function such as sum or maximum to reduce the problem to the univariate case. Closer to our aim in this paper is the work of Kim, Sun and Tsutakawa (2001) who use a Bayesian approach in a bivariate Poisson distribution to improve estimation of mortality rates in an autoregressive model.

A simple and effective approach to multivariate hotspot detection applies the univariate

ULS to each variable in the data set and identifies the univariate hotspot. Multivariate hotspots are those connected cells that appear in the intersection of the univariate hotspots of all variables. We will refer to this strategy as the Intersection method. The Joint method looks for zones Z by maximizing the likelihood of bivariate Binomial or Poisson models. The Intersection method works with the hotspots found by each variable and finds regions Z_X and Z_Y and defines $Z = Z_X \cap Z_Y$. Another approach to multivariate hotspot detection calls for the use of explanatory variables. Patil and Taillie (2004) characterize hotspots as regions of extreme departures from expectations. Hence, the size values A_a are proportional to model expectations and provide a link between a response variable and other explanatory variables. Regression techniques often provide a basis for adjusting the rates when a functional relationship is identified. To obtain hotspots based on all variables, the univariate ULS scan statistic is applied to the response variable and the adjusted sizes. The following approach provides univariate hotspots for each variable as well as hotspots that appear in the intersection set of the univariate hotspots. It directly accounts for the covariance of the variables and is based on a generalization of the univariate ULS scan statistic.

In this section we assume that (X, Y) is a random vector and extend the ULS scan statistic to a bivariate setting where we perform hotspot detection based on the joint effects of both variables. Similar to a univariate scan statistic, a region R of Euclidian space is partitioned into cells. For each cell a , observations are available in the form of quadruplets (X_a, Y_a, B_a, A_a) where X_a, Y_a and B_a are non-negative integers and A_a is a fixed and known constant. The bivariate ULS scan statistic exists naturally for the bivariate binomial and Poisson bivariate distributions. One can also explore the continuous case for bivariate distributions such as Log-normal and bivariate Exponential distributions.

4.1 Bivariate Binomial

Suppose $N_a = A_a$ people reside in cell a where each has two certain diseases with probabilities π_{xa} and π_{ya} . The variable X_a is a count of the number of people in cell a who have disease X . Similarly, Y_a counts the number of people in cell a who have disease Y . The variable B_a counts the number of people in cell a who have both diseases. One can also formulate an equivalent approach when a count of individuals who are disease-free is available for every cell. Consider Table I of bivariate Bernoulli distribution defined on cell a .

	Y=0	Y=1	Total
X=0	P_{00}	P_{01}	$1 - \pi_x$
X=1	P_{10}	P_{11}	π_x
Total	$1 - \pi_y$	π_y	1

The marginal distributions of X and Y are Bernoulli with parameters π_x and π_y . The marginal distributions of $X_a = \sum_{i=1}^{N_a} x_{ia}$ and $Y_a = \sum_{i=1}^{N_a} Y_{ia}$ are Binomial with N_a trials and

probabilities π_x and π_y , respectively. Let (X_{ia}, Y_{ia}) , for $i = 1, \dots, N_a$ be N_a independent copies from the bivariate Bernoulli distribution defined over cell a . The joint distribution of X_a and Y_a is bivariate Binomial with probability mass function

$$P(X_a, Y_a) = \sum_{k=0}^{\min(X_a, Y_a)} C_k P_{11}^k P_{10}^{X_a-k} P_{01}^{Y_a-k} P_{00}^{N_a-X_a-Y_a+k} \quad (2)$$

where $X_a, Y_a = 0, 1, \dots, N_a$ and $C_k = N_a! / (k!(X_a - k)!(Y_a - k)!(N_a - X_a - Y_a + k)!)$. In general, the parameters P_{11} , P_{10} , P_{01} and P_{00} depend on cell a . This dependence is assumed and suppressed for notational simplicity.

The observed counts X_a , Y_a and B_a are sufficient to estimate π_x , π_y and P_{11} . Note that if (X_a, Y_a) has a bivariate Binomial distribution with parameters $(P_{10}, P_{01}, P_{11}; N_a)$, then the correlation coefficient $\rho = \text{Corr}(X, Y)$ equals $(P_{11} - \pi_x \pi_y) / \sqrt{\pi_x(1 - \pi_x)\pi_y(1 - \pi_y)}$. Here, P_{11} is the probability that the two diseases occur together. Note that it is possible for one of the counts, say Y , to account for absence of a certain condition (disease), which may accompany X . In this case, the two disease counts are negatively correlated. In such cases the joint hot spot analysis is in fact a hot/cold spot analysis as we look for low values of one variable and high values of another.

We assume that the responses (X_a, Y_a) are independent of the counts in other cells and that spatial variability accounts for cell to cell variations in the parameters of the binomial model. In joint hotspot analysis, we scan the cells in region R and look for zones or neighborhood of connected cells with elevated responses relative to the rest of the region. Elevated responses are measured in terms of large values of the intensity function $G_a = (G_{X_a}, G_{Y_a})$ where $G_{X_a} = X_a/N_a$ and $G_{Y_a} = Y_a/N_a$. Under the null hypothesis of no joint hotspots, we state $H_0 : \pi_{xa} = \pi_x$ is the same for all cells a in R (no hotspots with respect to disease X), $\pi_{ya} = \pi_y$ is the same for all cells a in R (no hotspots with respect to disease Y), and that P_{11} is specified. Clearly, specifying the marginals, π_x and π_y , do not completely specify the distribution under the null hypothesis of no joint hotspots. We also need to specify P_{11} ; e.g. the probability of an individual with both diseases. We will study H_0 under different values of P_{11} . Note that when P_{11} is specified apriori (by specifying a correlation coefficient, for example) one does not need the individual counts B_a for each cell a , and only the pairs (X_a, Y_a) are used (see section 5). We first assume that the variables are independent; hence, $P_{11} = \pi_x \pi_y$ and study the hotspots obtained under independence. We will also assume that ρ_0 is the observed value of the sample correlation coefficient; i.e. $P_{11} = \rho_0 \sqrt{\pi_x(1 - \pi_x)\pi_y(1 - \pi_y)} + \pi_x \pi_y$. A third possibility is to set ρ and hence P_{11} at a fixed high (low) value. Using these values, one can study the sensitivity of the hotspots obtained and compare to the independence case.

The alternative hypothesis states that H_1 : there is a connected region of R (zone Z) and parameter values $0 < \pi_{x0} < \pi_{x1} < 1$ and $0 < \pi_{y0} < \pi_{y1} < 1$ such that for all cells a and

$$\pi_{x1} > \pi_{x0},$$

$$\pi_x = \begin{cases} \pi_{x1}, & \text{if } a \in Z, \\ \pi_{x0} & \text{if } a \in R - Z. \end{cases} \quad (3)$$

and

$$\pi_y = \begin{cases} \pi_{y1}, & \text{if } a \in Z, \\ \pi_{y0} & \text{if } a \in R - Z. \end{cases} \quad (4)$$

The alternative hypothesis is not completely specified either. We must specify the value of P_{11} under the alternative. Here, we will follow the same strategy as under the null hypothesis. Similar to the univariate approach, the zone Z in H_1 is a model parameter. The full model, $H_0 \cup H_1$ involves six unknown parameters $Z, \pi_{x0}, \pi_{x1}, \pi_{y0}, \pi_{y1}$, and P_{11} with $Z \in \Omega$, $\pi_{x1} > \pi_{x0}$ and $\pi_{y1} > \pi_{y0}$. For a given connected region, zone Z , the maximum likelihood estimates (MLE) of $(\pi_{x0}, \pi_{x1}, \pi_{y0}, \pi_{y1})$ and P_{11} can be stated explicitly. The profile likelihood for zone Z is $L(Z) = \max L(Z, \hat{\pi}_{x0}, \hat{\pi}_{x1}, \hat{\pi}_{y0}, \hat{\pi}_{y1}, P_{11})$ over $(\pi_{x0}, \pi_{x1}, \pi_{y0}, \pi_{y1}, P_{11})$. When P_{11} is not specified it can be replaced with \hat{P}_{11} . Following the search strategy of the univariate scan statistic, we will replace the full parameter space by a subset $\Omega_{ULS} \subset \Omega$, which is computationally feasible. The bivariate case is also adaptive in the sense that Ω_{ULS} is completely data-driven.

The rates define a piece-wise constant surface over the tessellation. This surface is 3-dimensional for each rate and 4-dimensional when both rates are considered. One can generalize the exceedance approach of defining the ULS to the multivariate setting. We may define the multivariate level $\vec{G} = (g, g, \dots, g) = g\vec{1}$ where $\vec{1} = (1, 1, \dots, 1)$ and multivariate exceedance $\vec{G}_a \geq \vec{g}$. In the bivariate case, multivariate exceedance is defined in terms of $(G_{xa} \geq g, G_{ya} \geq g)$. Thus, the multivariate ULS: $U_{\vec{g}} = \{a : \vec{G}_a \geq \vec{g}\}$. Similarly, we can define multivariate exceedance in terms of the levels of the norm $\sqrt{G_x^2 + G_y^2}$, $G_x + G_y$, $\max(G_x, G_y)$, $G_x G_y$, among others. This function is defined for all cells of R and over the vertices of the associated abstract graph. This function has a finite number of values (levels) in the tessellation and each level g determines an upper level set. The upper level sets that are identified do not need to be connected; hence, the reduced parameter list Ω_{ULS} consists of all connected components of all possible upper level sets. As in the univariate case, the size of Ω_{ULS} is less than the number of cells in the region R or the vertices in the abstract tree. The nodes of the ULS tree are candidate zones $Z \in \Omega_{ULS}$.

Let $x_T = \sum_a X_a$, the total number of cases of disease X and $y_T = \sum_a Y_a$, the total number of cases of disease Y . We also denote the total number of cases when both diseases are present with $B_T = \sum_a B_a$. When P_{11} is not pre-specified one estimates it with $\hat{P}_{11} = \sum_a B_a / n_T$.

Under H_0 , we estimate the rate of disease X and Y with $\hat{\pi}_x = x_T/n_T$ and $\hat{\pi}_y = y_T/n_T$, respectively. Note that $L(Null) = \prod_{a \in R} P(X_a, Y_a)$, where π_x, π_y and P_{11} are estimated with $\hat{\pi}_x, \hat{\pi}_y$ and \hat{P}_{11} . Under the alternative, $L(Z) = \prod_{a \in Z} P(X_a, Y_a) \prod_{a \notin Z} P(X_a, Y_a)$. Here, for $a \in Z$, we estimate π_x and π_y with $\hat{\pi}_{xz} = x_z/n_z$ and $\hat{\pi}_{yz} = y_z/n_z$, respectively, where $x_z = \sum_{a \in Z} X_a$, $y_z = \sum_{a \in Z} Y_a$ and $n_z = \sum_{a \in Z} N_a$. Similarly, for $a \notin Z$, we estimate π_x with $\hat{\pi}_{xz'} = X_{z'}/N_{z'}$ where $X_{z'} = \sum_{a \notin Z} X_a$ and $N_{z'} = \sum_{a \notin Z} N_a$. The likelihood ratio test of H_0 is $LR = L(Null)/L(Z)$.

To obtain the null distribution of LR one can use conditional simulation to generate samples from the bivariate binomial or the Poisson distribution. One can condition on the sufficient statistics under the null hypothesis to remove the unknown parameters from the null model. In the univariate case, the resulting distributions are hypergeometric and multinomial, respectively. One can easily generate samples from these distributions to find the sampling distribution of LR . However, in the bivariate case the resulting distributions do not take on a recognizable form and one must directly simulate an urn model to simulate from the distributions.

4.2 Bivariate Poisson Distribution

Campbell (1938) considers the following bivariate Poisson distribution. Let W_1, W_2 and W_{12} be independent Poisson random variables with means ϑ_x, ϑ_y and ϑ_B . Let $X = W_1 + W_{12}$ and $Y = W_2 + W_{12}$ with the joint probability mass function

$$P(X = X_a, Y = Y_a) = \exp -(\vartheta_x + \vartheta_y + \vartheta_B) \sum_{i=0}^{\min(X_a, Y_a)} \frac{\vartheta_x^{X_a-i} \vartheta_y^{Y_a-i} \vartheta_B^i}{(X_a - i)! (Y_a - i)! i!} \quad (5)$$

Hamdan and Al-Bayyati (1969) obtain (5) from (2) when P_{10}, P_{01} and P_{11} tend to zero as $N_a \rightarrow \infty$ in such a way that $N_a \pi_x \rightarrow \vartheta_x$, $N_a \pi_y \rightarrow \vartheta_y$ and $N_a P_{11} \rightarrow \vartheta_B$. The marginal distributions of X and Y are Poisson with means $\vartheta_x + \vartheta_B$ and $\vartheta_y + \vartheta_B$, respectively, and the correlation coefficient is $\rho = Corr(X, Y) = \frac{\vartheta_B}{\sqrt{(\vartheta_x + \vartheta_B)(\vartheta_y + \vartheta_B)}}$ with a maximum value of $\frac{\vartheta_B}{\vartheta_B + \min(\vartheta_x, \vartheta_y)}$ (Johnson, Kotz and Balakrishnan, 1997). Given k observations (x, y) from (5), Holgate (1964) shows that the MLE of ϑ_x and ϑ_y can be obtained from $\bar{x} - \hat{\vartheta}_B \bar{R} = \hat{\vartheta}_x$ and $\bar{y} - \hat{\vartheta}_B \bar{R} = \hat{\vartheta}_y$ where $\bar{R} = \sum_{a \in R} P(X_a - 1, Y_a - 1) / P(X_a, Y_a)$ and $\vartheta_B = Cov(X, Y)$ is estimated using the sample covariance $\hat{\vartheta}_B = \frac{1}{k-1} \sum_{a \in R} (X_i - \bar{x})(Y_i - \bar{y})$.

As in the binomial case, we assume $N_a = A_a$ people reside in cell a whose rates of diseases X and Y are given by Poisson means. For each cell a observations (X_a, Y_a, B_a, A_a) are available. In general, the parameters ϑ_x, ϑ_y , and ϑ_B depend on cell a . For the Poisson model, under the null hypothesis of no joint hotspots, we state: $H_0 : \vartheta_{xa} = \vartheta_x$ (no hotspots with respect to X) and $\vartheta_{ya} = \vartheta_y$ (no hotspots with respect to Y) for all cells a in R .

We also need to specify the covariance ϑ_B . Three cases are considered, 1) $\vartheta_B = 0$ under independence, 2) estimate value of the covariance by its MLE $\hat{\vartheta}_B$ and 3) specify a high (low) value for ϑ_B to study the sensitivity of hotspots and compare against the independence case. For a zone Z , the MLE of $(\vartheta_x, \vartheta_y, \vartheta_B)$ is obtained following Holgate (1964) and the profile likelihood for the zone Z is $L(Z) = \max L(Z, \hat{\vartheta}_{x0}, \hat{\vartheta}_{x1}, \hat{\vartheta}_{y0}, \hat{\vartheta}_{y1}, \hat{\vartheta}_B)$. Under the null hypothesis $L(Null) = \prod_{a \in R} P(X_a, Y_a)$ where $(\hat{\vartheta}_x, \hat{\vartheta}_y, \hat{\vartheta}_B)$ estimate the parameters. Under the alternative the likelihood based on (5) is evaluated for $a \in Z$ and $\hat{\vartheta}_x = x_z/n_z$, $\hat{\vartheta}_y = y_z/n_z$, and the sample covariance $\hat{\vartheta}_B$.

4.3 Other Test Statistics

One can also consider other test statistics than the likelihood ratio test for detection of hotspots. Suppose $(\vec{X}_1, \dots, \vec{X}_k)$ be a sample from a p -dimensional distribution with mean vector $\vec{\theta}$. For the binomial distribution we have (X_a, Y_a) and B_a available at each of the k cells in R , each with mean $\vec{\theta} = (\pi_x, \pi_y)$ and P_{11} . The univariate scan statistic examines candidate zones corresponding to high response rate relative to a set of prior expected responses. Instead of estimating the expected responses one can measure how far the observed rates deviate from a baseline rate θ_0 , which is a pre-specified value of θ . To test $H_0 : \vec{\theta} = \vec{\theta}_0$ note that H_0 is true whenever $\vec{\alpha}'\vec{\theta} = \vec{\alpha}'\vec{\theta}_0$ is true for all $\vec{\alpha}$. For fixed $\vec{\alpha}$, we can test H_α , the hypothesis that $\vec{\alpha}'\vec{\theta} = \vec{\alpha}'\vec{\theta}_0$ against H_1 , the alternative that $\vec{\alpha}'\vec{\theta} \neq \vec{\alpha}'\vec{\theta}_0$. Let $w_i = \vec{\alpha}'\vec{G}_i$ be k independent copies from random variable W where $G_i = X_i/N_i$ denotes the p observed rates in cell i for $i = 1, \dots, k$. Hence, H_0 is equivalent to $E(W) = \vec{\alpha}'\vec{\theta}_0$. When $p = 1$, one may use $t_\alpha = \sqrt{n_z}(\bar{w} - \theta_0)/s_w$ where \bar{w} is the mean and s_w is the standard deviation of some candidate zone Z . As noted by a referee, w_i have heterogeneous variances; hence, t_α is an inefficient estimator. One should weigh w_i ; that is, replace \bar{w} with $[\sum \frac{w_i}{var(w_i)}]/[\sum 1/var(w_i)]$ and base a statistic on the weighted version. Roy's (1953) union-intersection principle rejects H_0 for large values of

$$T_z = \max |t_\alpha| = \sqrt{n_z}[(\vec{w} - \vec{\theta}_0)'s^{-1}(\vec{w} - \vec{\theta}_0)]^{\frac{1}{2}}$$

over all $\vec{\alpha}$ where \vec{w} and s are the sample mean and covariance matrix of zone Z . We will find $T = \max(T_z)$ as the top hotspot for all candidate zones Z . Siotani (1959) points out that the sampling distribution of this type of statistic is extremely difficult to find analytically. However, the sampling distribution of this statistic can easily be obtained by simulating observations from the binomial or Poisson distribution. When the responses have a multivariate normal distribution, T_z has a Hotelling- T^2 statistic (Anderson, 1984; Roy, 1957) and is a function of the likelihood ratio test. It is well known that T^2 statistic is uniformly most powerful in the class of invariant tests. This statistic is designed to detect shift alternatives of the form $\vec{\theta} = \vec{\theta}_0 + \vec{\delta}$.

5 Sensitivity Analysis

How sensitive are the joint hotspots to the degree of association between X and Y ? We do not expect to see common hotspots when X and Y are independent whereas as the strength of association between the variables increases, we expect to see many more common hotspots. In some cases information on B_a , the number of individuals with both diseases in cell a may not be available apriori. We would like to impose a new correlation between the two variables in order to compare the joint hotspots to the ones obtained using the Intersection method or under the assumption of independence. Consider the bivariate binomial model and pairs of random observations (X_a, Y_a) , where X and Y have marginal binomial distributions, with a given degree of association.

At each cell a in R , we simulate a bivariate binomial random vector with parameters $\hat{\pi}_x = x_T/n_T$, $\hat{\pi}_y = y_T/n_T$, and $\hat{P}_{11} = \rho_0\sqrt{\hat{\pi}_x(1-\hat{\pi}_x)\hat{\pi}_y(1-\hat{\pi}_y)} + \hat{\pi}_x\hat{\pi}_y$. The resulting data set will be used to obtain the new hotspots with the correlation, ρ . The generated sample will exhibit marginal hotspots that are similar to the ones obtained from the original data. The joint hotspots will reflect the effects of the new degree of association on the data. We first assume that the variables are independent; hence, $P_{11} = \pi_x\pi_y$ or $\rho = 0$ and study the hotspots obtained under independence. We will also assume that ρ_0 is the observed value of the sample correlation coefficient. A third possibility is to set ρ and hence P_{11} at a fixed high (low) value. Using these values, one can study the sensitivity of the hotspots obtained and compare to the independence case.

Using computer simulation as described, one can construct a table that tells us how large ρ has to be for the intersection hotspot to be valuable. One can define the ratio of the number of cells in the intersection hotspots to the total number of cells in the univariate hotspots as a measure of effectiveness of the joint and intersection hotspot methods. The value of this ratio approaches 1 as ρ tends to 1 and approaches 0 as ρ tends to 0.

6 Application

Also called hot addresses (Eck and Weisburd, 1995; Sherman, Gartin and Buerger, 1989), crime hotspots are concentrations of individual events that suggest a series of related crimes (Eck, Chainey, Cameron, Leitner and Wilson, 2005). Similar to disease counts, crime rates are not uniformly distributed across the tessellation. Crime is usually more prevalent in some areas while largely absent in others. Allocation of resources is usually based on where the demand for law enforcement is highest. The uniform crime reporting program (ICPSR, 2004) provides data collected at the county-level for all states and several offenses, including murder, rape, robbery, aggravated assault, burglary, larceny, auto theft, among others. Robbery is defined as taking of personal property in the possession or immediate presence

of another by the use of violence or intimidation. Residential burglary is the act of breaking into a house to commit theft or other felony.

Figure 6 shows the top 5 significant hotspots for data on Burglary. Significant hotspots are marked with X. Figure 7 shows the top three significant hotspots for the data on Robbery. Figure 8 shows the top significant hotspots for the intersection of Burglary and Robbery.

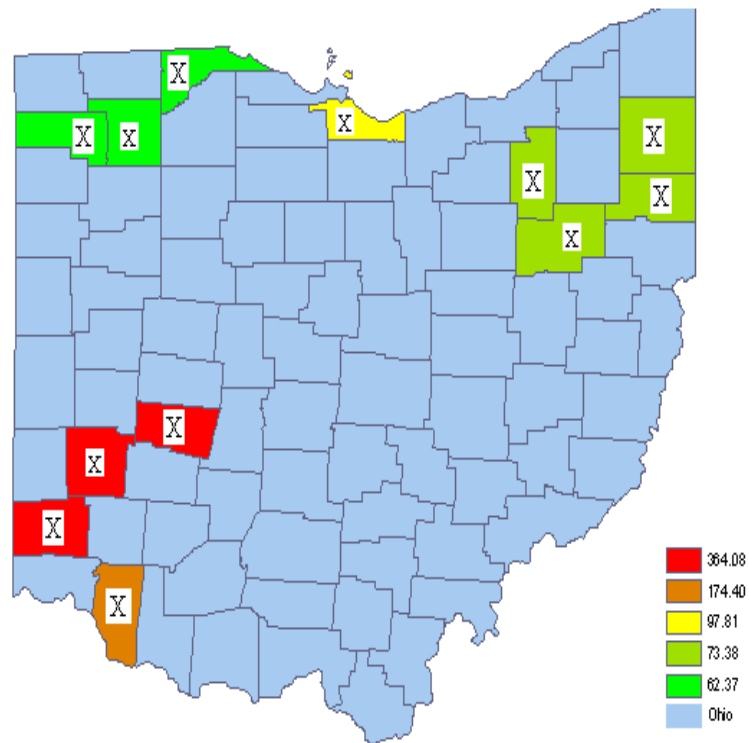


Figure 6: The top five hotspots of Burglary (marked with X) in counties of the state of Ohio are significant at 0.001 level. Data source is ICPSR-4009, Uniform Crime Reporting Program Data [United States]: County-Level Detailed Arrest and Offense Data, 2002.

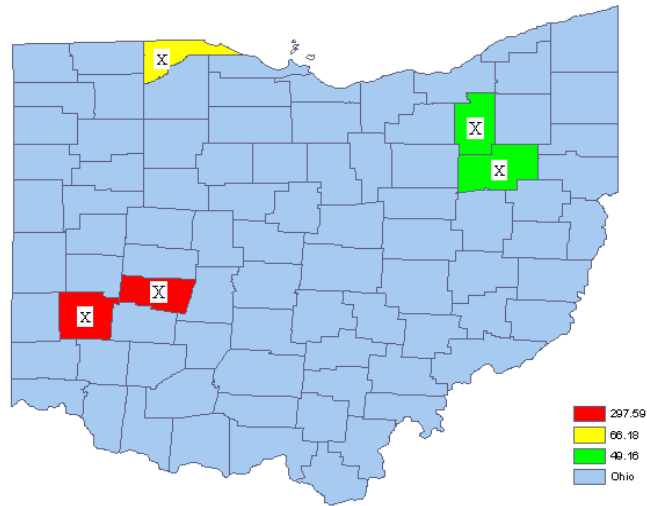


Figure 7: The top three hotspots of Robbery (marked with X) in counties of the state of Ohio are significant at 0.001 level. Data source is ICPSR-4009, 2002.

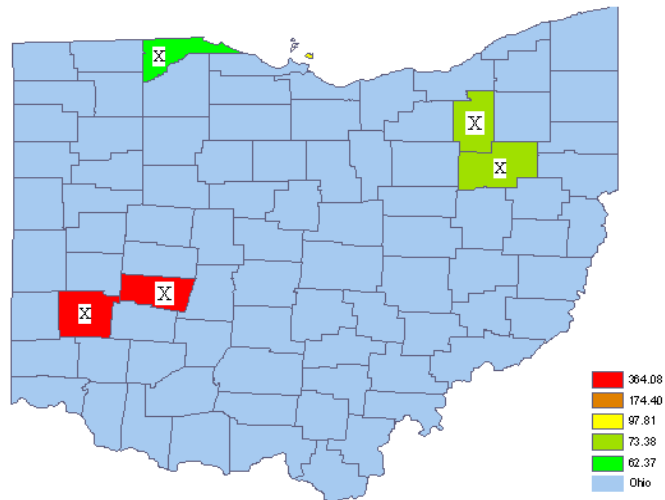


Figure 8. The top significant hotspots (marked with X) at 0.001 level obtained by the Intersection method for Burglary and Robbery in counties of the state of Ohio, 2002.

7 Summary and Concluding Remarks

We have discussed the need for detection of hotspots. Any hotspot detection method is composed of three main components: identification of candidate hotspots, evaluation of their statistical significance and assessment of covariates. We discuss the Upper Level Scan statistic (Patil and Taillie, 2004) for detection of hotspots, its theory and design implementation. The heart of the ULS scan statistic is the ULS tree. We provide the ULS-Hotspot algorithm that obtains the rates, maintains a list of connected components at each level of the rates and yields the ULS tree. The tree is grown in the immediate successor list, which provides a computationally efficient method for likelihood evaluation and storage. An example shows how the zones are formed and the likelihood function is developed for each candidate zone. We note that the ULS scan statistic is univariate in its development and extend it to the bivariate case. The general theory of bivariate hotspot detection is explained, including the bivariate binomial model, the multivariate exceedance approach, the bivariate Poisson distribution. We propose the Intersection method that is simple to implement, using a univariate hotspot detection method. Along with the likelihood ratio test, we examine a quadratic form of the rates to identify deviations from a baseline rate.

We ask how sensitive the joint hotspots are to the degree of association between the variables and propose two methods for sensitivity analysis. The simple method is simulation based and easy to implement. We examine the mapping of crime hotspots. In crime analysis, hotspots are defined as concentrations of individual events that suggest a series of related crimes. Allocation of resources is usually based on where the demand for law enforcement is highest and crime hotspots play an important tool of the crime analyst. We study the univariate hotspots and the Intersection method for Robbery and Burglary in state of Ohio and obtain their univariate and intersection hotspots.

References

- [1] Anderson, T. W. (1984). *An Introduction to Multivariate Analysis*. Second Edition. John Wiley & Sons. New York.
- [2] Burkom, H. S. (2003). Biosurveillance applying scan statistic with multiple, disparate data sources. *Journal of Urban Health*, 80, 57-65.
- [3] Campbell, J. T. (1938). The Poisson correlation function. *Proceedings of the Edinburgh Mathematical Society (Series 2)*, 4, 18-26.
- [4] Dwass, M. (1957). Modified randomization tests for nonparametric hypothesis. *Annals of Mathematical Statistics*, 28, 181-187.
- [5] Duczmal, L., Kulldorff, M. and Mostashari, F. (2005). Power evaluation of the spatial scan statistic for multiple data streams. Submitted for publication.
- [6] Eck, J. E. and Weisburd, D. (1995). Crime places in crime theory. In J. E. Eck and D. Weisburd (eds.) *Crime Places*, Vol. 4, 1-33. Monsey, NY. Crime Justice Press.
- [7] Eck, J. E., Chainey, S., Cameron, J. G., Leitner, M. and Wilson, R. E. (2005). *Mapping Crime: understanding hotspots*. National Institute of Justice (<http://www.opj.usdoj.gov/nij>).
- [8] Hamdan, M. A., and Al-Bayyati, H. A. (1969). A note on the bivariate Poisson distribution. *The American Statistician*, 23, No. 4, 32-33.
- [9] Holgate, P. (1966). Estimation of the bivariate Poisson distribution. *Biometrika*, 51, 241-245.
- [10] Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions*. John Wiley & Sons. New York.
- [11] ICPSR (2004). U.S. Department of Justice, Federal Bureau of Investigation. Uniform Crime Reporting Program Data: County-Level Detailed Arrest and Offense data. <http://www.icpsr.umich.edu/>.
- [12] Kim, H., Sun, D. and Tsutakawa, R. K. (2001). A bivariate method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association*, Vol. 96, No. 456, 1506-1521.
- [13] Kulldorff, M. and Nagarwalla, N. (1995). Spatial disease clusters: Detection and inference. *Statistics in Medicine*, 14, 799-810.
- [14] Kulldorff, M. (1997). A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26, 1481-1496.

- [15] Li, H., Calder, C. A., and Cressie N. (2005). Beyond Morans I: Testing for Spatial Dependence based on the SAR model. Preprint No. 763, Department of Statistics, the Ohio State University.
- [16] Moran, P. A. P. (1950). Notes on continuous stochastic phenomena.. *Biometrika*, 37, 17-23.
- [17] Patil, G. P. and Taillie, C. (2004). Upper level set statistic for detecting arbitrarily shaped hotspots. *Environmental and Ecological Statistics* 11, 183-197.
- [18] Patil, G. P., Modarres, R. and Patakar, P. (2005). The ULS software, version 1.0. Center for Statistical Ecology and Environmental Statistics. Department of Statistics, Pennsylvania State University.
- [19] Roy, S. N. (1953). On a heuristic method of test construction and its use in multivariate analysis. *Annals of Mathematical Statistics*, 24, 220-238.
- [20] Roy, S. N. (1957). *Some Aspects of Multivariate Analysis*. John Wiley & Sons. New York.
- [21] Sherman, L. W., Gartin, P. R. and Buerger, M E. (1989). Hotspots of predatory crime: routine activities and criminology of place. *Criminology*, V. 27, 1, 27-55.
- [22] Siotani, M. (1959). The extreme value of the generalized distance of the individual points in the multivariate normal sample. *Ann. Institute Statist. Math.*, Tokyo, 10, 183-208.
- [23] Wong, W. K., Moore, A., Cooper G., Wagner, M. (2003). WSARE: What's strange about recent events? *Journal of Urban Health*, 80, 66-75.