Available online at www.sciencedirect.com
ScienceDirect
Journal of Statistical Planning and Inference 137 (2007) 3643–3654
journal of statistical planning and inference
www.elsevier.com/locate/jspi

Hotspot detection with bivariate data

Reza Modarres[a,1], G.P. Patil[b,*,2]

[a]*Department of Statistics, The George Washington University, Washington, DC 20052, USA*
[b]*Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA*

Available online 7 April 2007

Abstract

The upper level set (ULS) scan statistic, 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 method. 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

Keywords: Hotspots; Geoinformatic surveillance; Spatial scan statistic; Upper level set scan statistic; Bivariate; Crime mapping

1. Introduction

Geoinformatic surveillance for the 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 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 the 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)

* Corressponding author. Tel.: +1 814 865 9442; fax: +1 814 865 1278.
E-mail addresses: Reza@gwu.edu (R. Modarres), gpp@stat.psu.edu (G.P. Patil).

[1] This work was completed while the author was on sabbatical at the Center for Statistical Ecology and Environmental Statistics of the Pennsylvania State University.

[2] This 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.

0378-3758/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2007.03.040

scan statistic (Patil and Taillie, 2004). SaTScan is available on the Internet and the ULS scan is under development (Patil et al., 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 log-normal 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 by Ω (see Fig. 1).

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 50% 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: there is a zone Z such that for $P_1 > P_0$,

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

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

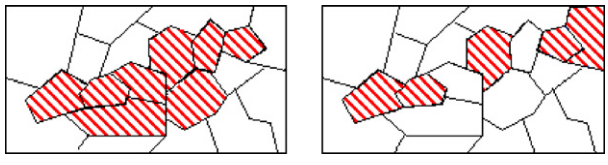

Fig. 1. A tessellated region. The collection of shaded cells on the left is connected, hence a zone Z in Ω. The collection on the right is not connected.