ARTICLE IN PRESS

Spatial Statistics ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Spatial Statistics

journal homepage: www.elsevier.com/locate/spasta

A distribution-free spatial scan statistic for marked point processes

Lionel Cucala *

Institut de Mathématiques et Modélisation de Montpellier, France

ARTICLE INFO

Article history:
Received 4 October 2013
Accepted 25 March 2014
Available online xxxx

Keywords:
Spatial statistics
Scan statistics
Cluster detection
Non-parametric methods

ABSTRACT

A new spatial scan statistic is proposed for identifying clusters in marked point processes. Contrary to existing methods, it does not rely on a likelihood ratio and thus is completely distribution-free. It applies whatever the nature of the marks: binary, discrete or continuous. This spatial scan test seems to be very powerful against any arbitrarily-distributed cluster alternative. I apply this method first to a classical epidemiological dataset and then to the spatial distribution of incomes in France.

1. Introduction

Cluster detection has become a very fruitful research subject since the earlier work of Naus (1963): a thorough review of the proposed methods, which have been applied to many different fields of application, is given by Glaz et al. (2001).

Most of the cluster detection methods are designed for count data, i.e. point processes made of the random coordinates of n events observed in S, a bounded subset of $\mathbb{R}^d$: the goal is to identify, if they exist, the areas in which the concentration of events is abnormally high. Since the article by Cressie (1977), the scan statistic denotes the maximal concentration observed on a collection of potential clusters. Originally, the size of all the potential clusters had to be the same, so that the scan statistic was just the maximum number of events in a window of size d, d being fixed a priori. This major drawback vanished when Kulldorff (1997) introduced the scan statistic based on generalized likelihood-ratio (GLR) in a Poisson model, which allows to compare the concentration in windows having different sizes. In the same article, the Bernoulli model scan statistic is defined to analyse point processes with

* Tel.: +33 467143956.
E-mail address: lcucala@math.univ-montp2.fr.

http://dx.doi.org/10.1016/j.spasta.2014.03.004
2211-6753/

Please cite this article in press as: Cucala, L., A distribution-free spatial scan statistic for marked point processes. Spatial Statistics (2014), http://dx.doi.org/10.1016/j.spasta.2014.03.004

binary marks, such as case/control data: if the marks of the cases are 1 and those of the controls are 0, the goal is to identify the areas in which the marks are significantly higher, i.e. the areas where there are significantly more cases, taking into account the number of controls. Later on, Kulldorff et al. (2009) introduced the Gaussian model scan statistic which allows to analyse point processes with continuous marks.

For the analysis of point processes on the line, Cucala (2008) suggested that the use of a non-parametric concentration index may be more powerful to detect cluster presence than the ones based on likelihood-ratio tests, such as the one introduced by Nagarwalla (1996). Thus, in order to analyse spatial marked point processes, I may look for a concentration index only relying on the distribution-free (DF) null hypothesis H_0: "the marks are realizations of independent and identically distributed random variables". Very recently, Cucala (in press) defined a Mann–Whitney scan statistic based on the same hypothesis but, as it relies only on the ranks of the marks, it is only suitable for continuous marks.

In this article I introduce a scan statistic for any kind of marked point processes. Section 2 describes the scan statistic and its computational aspects in the framework of marked point processes. The scan statistic is then applied to real and simulated datasets in Section 3. The paper is concluded with a discussion.

2. A distribution-free scan statistic

Let $\{(x_i, s_i), i = 1, \ldots, n\}$ denote the realization of a marked point process, where $s_i \in S$ is the spatial location of the event and $x_i \in \mathbb{R}$ its associated mark. The area $S \subset \mathbb{R}^d$ is the observation domain and the spatial locations are usually bidimensional ($d = 2$). Our goal is to detect the spatial area $Z \subset S$ in which the marks are significantly different (higher or lower) than elsewhere.

Most of the spatial cluster detection methods consist in maximizing (or minimizing) a concentration index in a collection of potential clusters. Thus the two questions to answer are: how to choose the potential clusters and which concentration index should be used?

Concerning the potential clusters, I will focus on circular clusters, such as (Kulldorff, 1997). The set of potential clusters, denoted by $\mathcal{D}$, is the set of discs (or balls if $d = 3$) centred on a location and passing through another one:

$$\mathcal{D} = \{D_{i,j}, 1 \leq i \leq n, 1 \leq j \leq n\}$$

where $D_{i,j}$ is the disc (or the ball) centred on s_i and passing through s_j. Since the disc may have null radius (if $i = j$), the number of potential clusters is n^2.

As said in the Introduction, most of the concentration indices are based on generalized likelihood ratio and the choice of the underlying distribution, as mentioned by Cucala (2008) in the unidimensional context, may lead to very different results. The concentration index I introduce here is only based on the DF null hypothesis H_0 given in the Introduction, without specifying any distribution. In the works by Kulldorff (1997) or Kulldorff et al. (2009), the null hypothesis is the same but the distribution of the marks is mentioned (Bernoulli, Poisson or Gaussian). All these studies rely on the assumption that the variances of the marks are all equal. A work by Huang et al. (2009) extends the Gaussian scan statistic to the marks with unequal variances, for example when the marks are means computed on different population sizes and the individual data are not available. Thus I first give the details of my distribution-free method when variances are assumed to be equal, and then an heteroskedastic version taking into account unequal variances.

2.1. The homoskedastic version

From now on, I assume H_0 is true. Let $X_1, \ldots, X_n$ denote the i.i.d. random variables associated to the marks. I will assume that the distribution of the marks has a second moment so that

$$\mathbf{E}(X_i) = \mu \quad \text{and} \quad \mathbf{V}(X_i) = \sigma^2 \quad \text{for all } i$$

and

$$Cov(X_i, X_j) = 0 \quad \text{if } i \neq j.$$

The common expectation μ and variance σ^2 of the X_i's are unknown.

Please cite this article in press as: Cucala, L., A distribution-free spatial scan statistic for marked point processes. Spatial Statistics (2014), http://dx.doi.org/10.1016/j.spasta.2014.03.004