Spatial Statistics 21 (2017) 66–74

Contents lists available at ScienceDirect

Spatial Statistics

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

A multivariate Gaussian scan statistic for spatial data

Lionel Cucala [a,*], Michaël Genin [b], Caroline Lanier [c], Florent Occelli [c]

[a] Institut Montpelliérain Alexander Grothendieck, Université de Montpellier, France
[b] Santé Publique: épidémiologie et qualité des soins EA 2694, Université Lille, France
[c] IMPact de l'Environnement Chimique sur la Santé humaine EA 4483, CHU Lille, Institut Pasteur de Lille, Université Lille 2, France

ARTICLE INFO

Article history:
Received 8 March 2017
Accepted 6 June 2017
Available online 15 June 2017

Keywords:
Spatial statistics
Scan statistics
Cluster detection

ABSTRACT

A new spatial scan statistic is proposed for multivariate data indexed in space. Such as many other scan methods, it relies on a generalized likelihood ratio but it also takes into account the correlations between variables. This spatial scan test seems to be more powerful than the independent version, whatever the level of correlation between variables. We apply this method to a data set recording the levels of pollutant metals in the area of Lille, 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. (2009).

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 in a Poisson model, which allows to compare the concentration in windows having different sizes. In

* Corresponding author.
E-mail address: lionel.cucala@umontpellier.fr (L. Cucala).

http://dx.doi.org/10.1016/j.spasta.2017.06.001
2211-6753/

the same article, the Bernoulli model scan statistic is defined to analyse point processes with 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.

Sometimes, such as in environmental surveillance, numerous continuous variables have to be analysed in the same time. A multivariate scan statistic combining different univariate scan statistics has been proposed by Kulldorff et al. (2007) and presents the great advantage that both count data and continuous data can be simultaneously analysed. However, this scan statistic does not take into account the covariances between different variables.

In this paper we develop a scan statistic for multivariate continuous data that is based on the Gaussian model and takes into account the covariances between variables. Under the null hypothesis, all observations come from the same distribution. Under the alternative hypothesis, there is one cluster location where the observations have a different mean vector than outside that cluster. A key feature of the method is that the statistical inference is still valid even if the true distribution is not Gaussian, assuring that the correct alpha level is maintained. This is accomplished by evaluating the statistical significance of clusters through a permutation-based Monte Carlo hypothesis testing procedure. Section 2 describes the scan statistic and its computational aspects. The scan statistic is then applied to real and simulated datasets in Section 3. The paper is concluded with a discussion.

2. A scan statistic for multivariate data

Let $X^1, \ldots, X^p$ denote a collection of p variables which are measured in n different spatial locations $s_1, \ldots, s_n$ included in D. The area $D \subset \mathbb{R}^d$ is the observation domain and the spatial locations are usually bidimensional ($d = 2$). The measure of variable X^j in location s_i is denoted by x_i^j and all measures are recorded in a $n \times p$ matrix

$$X = (x_i^j), \quad 1 \leq i \leq n, \quad 1 \leq j \leq p.$$

The $1 \times p$ vector containing all the measures in s_i, and corresponding to the ith row of matrix X, is denoted by X_i. Our goal is to detect the spatial area $Z \subset D$ in which the measures of the variables are significantly different (higher or lower) than elsewhere.

Most of the spatial cluster detection methods consist in maximizing a likelihood ratio in a collection of potential clusters. Thus the two questions to answer are: how to choose the potential clusters and which likelihood ratio should be used?

Concerning the potential clusters, we will focus on variable-size 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. Remark that many other possibilities, such as elliptic clusters (Kulldorff et al., 2006) or graph-based (Cucala et al., 2012), have been proposed.

As said in the Introduction, Kulldorff et al. (2009) introduced a Gaussian-based scan statistic to detect clusters when dealing with univariate continuous data. It relies on the likelihood ratio between two hypotheses: the marks are supposed to be normally-distributed and independent but the null hypothesis considers equal means and variances whereas the alternative hypothesis considers equal variances but different means inside and outside the potential cluster. Our method extends this procedure to the multivariate case.

The random vectors $X_1, \ldots, X_n$, i.e. the measures associated to the n different locations, are assumed to be independent: this is a very classical assumption when introducing likelihood-based scan statistics. The null hypothesis H_0, corresponding to the absence of any cluster in the data, is the following:

$$X_i \sim \mathcal{N}_p(\mu^*, \Sigma^*), \quad \forall i = 1, \ldots, n,$$