Spatial and Spatio-temporal Epidemiology 5 (2013) 39–50

Contents lists available at SciVerse ScienceDirect

Spatial and Spatio-temporal Epidemiology

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

# Spatial clusters in a global-dependence model

Tai-Chi Wang *, Ching-Syang Jack Yue

*Department of Statistics, National Chengchi University, No. 64, Sec. 2, ZhiNan Rd., Wenshan District, Taipei City 11605, Taipei, Taiwan, ROC*

ARTICLE INFO

*Article history:*
Received 7 October 2011
Revised 22 January 2013
Accepted 8 March 2013
Available online 13 April 2013

*Keywords:*
Local cluster
Spatial global dependence
Conditional autocorrelated regressive model
Spatial scan statistic
EM estimates
Generalized least square

ABSTRACT

Spatial data often possess multiple components, such as local clusters and global clustering, and these effects are not easy to be separated. In this study, we propose an approach to deal with the cases where both global clustering and local clusters exist simultaneously. The proposed method is a two-stage approach, estimating the autocorrelation by an EM algorithm and detecting the clusters by a generalized least square method. It reduces the influence of global dependence on detecting local clusters and has lower false alarms. Simulations and the sudden infant disease syndrome data of North Carolina are used to illustrate the difference between the proposed method and the spatial scan statistic.



## 1. Introduction

In spatial data analysis, one of the frequently discussed issues is the relationship between geographical locations, that is, the identification of spatial patterns. The particular interest is whether certain locations are significantly different from other locations in the aspect of statistical testings. Besag and Newell (1991) categorized such tests into two types: general tests and focused tests. For the general tests, it can be further categorized as global clustering and cluster detection tests. Kulldorff et al. (2006) gave a great amount of references of these tests.

Global clustering tests, such as Moran's I statistic and Geary's C statistic, are concerned with global clustering patterns. Global clustering patterns can be modeled by using spatial autoregressive models or conditional autoregressive models (Besag, 1974; Cressie, 1993). On the other hand, cluster detection tests are used to determine if some attributes of one or more subregions, such as incidence rates of disease, are unusually large, that is, to identify hot spots. Getis and Ord (1992) and Anselin (1995) discussed several statistics to test local dependence. In recent years, spatial cluster detection methods have been widely applied to many different fields.

However, the efficiency of cluster detection methods is often data-dependent. Among these methods, the spatial scan statistic (SaTScan) (Kulldorff and Nagarwalla, 1995) is perhaps the most popular and is considered to be quite effective in many instances. For example, Huang et al. (2008) had compared several cluster detection methods, such as circular and elliptic spatial scan statistics (SaTScan), flexibly shaped spatial scan statistics, Turnbull's cluster evaluation permutation procedure, local indicators of spatial association, and upper-level set scan statistics. They found that the SaTScan had the best performance in several synthetic cluster patterns. Although the past studies have shown that the SaTScan is quite effective in detecting spatial clusters (Kulldorff et al., 2003; Takahashi and Tango, 2006; Huang et al., 2008), few of these studies (and other cluster detection methods) discuss the performance of cluster detection in case of global spatial autocorrelation. Besides, it should be noted that the SaTScan relies on the Monte Carlo method to decide the significance of clusters, and the global spatial autocorrelation can distort

* Corresponding author. Tel.: +886 2 2939 3091x81144; fax: +886 2 8237 0079.

*E-mail address:* taichi@alumni.nccu.edu.tw (T.-C. Wang).

1877-5845/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.sste.2013.03.003

the Monte Carlo results. Intuitively, cluster detection can be expected to be more accurate when considering the global dependence.

It should be noted that using different cluster analysis methods, such as globally autoregressive models and cluster detection methods, can also produce different interpretations. For example, Cressie and Read (1989) discussed spatial autoregressive models on the sudden infant death syndrome (SIDS) data from 1974 to 1984 for the counties of North Carolina and concluded that the errors show some (nonsignificant) spatial dependent structure. On the other hand, Kulldorff (1997) identified several local clusters using the SaTScan for the same data but with different combinations of all years. In other words, a pattern of one geographical scale can be identified as another spatial pattern.

In identifying spatial cluster patterns, almost all cluster detection methods assume that the data are independent rather than dependent. However, the result of cluster detection may probably be affected by the local effects, as well as global dependence. Ord and Getis (1995) showed that "when global autocorrelation exists, local pockets are harder to detect." Ord and Getis (2001) said, "If existing tests are applied without regard to global autocorrelation structure, type I errors may abound." They provided the local $O$ statistic, which can accommodate spatial parameters identified from variograms and correlograms, to detect local clusters. However, their method did not consider that local clusters also affect the estimate of autocorrelation. Kulldorff (2006) depicted the difficulties of identifying these two patterns and gave a general framework for testing the spatial randomness. Lawson (2006) differentiated these two effects and gave more specific definitions of them. Although there are many articles describing the possibility of existing multiple patterns in spatial data, the solution to the model involved local clusters and global dependence is rarely mentioned. In this study, in addition to showing the difficulty of disentangling these two effects, we propose a method including both local clusters and global clustering.

In this paper, we propose a cluster detection approach that deals with global spatial dependence. Before introducing the proposed method, we will first review the conditional autoregressive spatial model (or the conditionally specified spatial Gaussian model) and the SaTScan in Section 2. In addition, we will evaluate the performance of the SaTScan in case of spatial autocorrelation. Then, the proposed approach is introduced, together with the EM estimates, the scanning procedure, and the Monte Carlo testing for handling both global dependence and clusters in Section 3. Simulations and an empirical study (the SIDS data of North Carolina) are used to evaluate the proposed methods in Sections 4 and 5. We will present comments and discussions on the proposed approach in the final section.

## 2. Spatial models and the SaTScan

We first introduce the concepts of global dependence and cluster detection. Regarding to global dependence, we will provide a brief introduction of spatial models. For a complete introduction to spatial autoregressive models, please refer to Chapter 6 of Cressie (1993). To identify spatial clusters, the concept of the SaTScan is briefly introduced, and a detailed discussion of which can be found in Kulldorff's work (Kulldorff, 1997). Furthermore, we will demonstrate that the performance of the SaTScan can be influenced by global dependence.

### 2.1. Conditional autocorrelated regressive model with Gaussian distribution

For the global dependence model, we will only use the conditional autocorrelated regressive (CAR) model in this study. The CAR model is considered on the basis of its popularity and can be used in spatial regression models. For a CAR model, $\{Z(s_i) : s_i \in D,\ \forall i \in \{1, 2, \ldots, n\}\}$ is defined as a spatial process, or a random process in the spatial domain in lattice $D$, and $s_i = (u_i, v_i)$ is the location of cell $i$, where $(u_i, v_i)$ are the coordinates. In practice, $s_i$ is often defined as the geographic center of cell $i$. Suppose the full conditional distribution of $Z(s_i)$ which follows a Gaussian distribution can be expressed as

$$f(z(s_i)|\{z(s_j) : j\neq i\}) = \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left[-\frac{\{z(s_i) - \theta_i(\{z(s_j) : j\neq i\})\}^2}{2\sigma_i^2}\right], \quad (1)$$

where $f$ denotes the conditional density function of $z(s_i)$, and $\theta_i$ and $\sigma_i^2$ are the conditional mean and variance respectively. The term "pairwise-only dependence" is defined as the condition of $\theta_i$ satisfying

$$\theta_i(\{z(s_j) : j\neq i\}) = \mu_i + \sum_{j\neq i} c_{ij}(z(s_j) - \mu_j). \quad (2)$$

Let us assume that the weight of neighborhood information, $c_{ij}$, is equal to $\rho \times w_{ij}$, where $w_{ij}$ is a known weight and $\rho$ is an unknown spatial dependent parameter. Along with the Hammersley–Clifford theorem, the joint distribution of $\mathbf{Z} \equiv (Z(s_1), \ldots, Z(s_n))^T$, the CAR model, can be established as $\mathbf{Z} \sim Gau(\mu, (I - \rho \times W)^{-1}M)$, where $\mu = (\mu_1, \ldots, \mu_n)^T$ is the mean vector, $W$ is an $n \times n$ matrix whose $(i,j)$ element is $w_{ij}$, $M \equiv diag(\sigma_1^2, \ldots, \sigma_n^2)$ is an $n \times n$ diagonal matrix, and $(I - \rho \times W)$ is necessarily symmetric and invertible (Besag, 1974; Cressie, 1993).

To estimate the unknown parameters in the CAR model, the maximum likelihood estimates (MLEs) are the most popular method. However, the MLEs do not have the closed forms and this reason could be an intractable problem when we need to do some further computations. Besag (1974) introduced the pseudo-likelihood to solve it and has been proved that the estimates are consistent. We will apply the pseudo-likelihood to obtain the EM estimates in case of clusters which will be discussed in Section 3.

Besides, the selection of a suitable weight function is a difficult problem. In this paper, we will use the "C" type weight function, which is a globally standardized function (the number of cells divided by sums over all number of neighbors in the study region), because it can maintain a