Journal of Statistical Planning and Inference 139 (2009) 136 – 142

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Testing spatial randomness based on empirical distribution function: A study on lattice data

Dejian Lai[a, b, *]

[a] Division of Biostatistics, School of Public Health, University of Texas, 1200 Herman Pressler, Suite 1006, Houston, TX 77030, USA
[b] Faculty of Statistics, Jiangxi University of Finance and Economics, Nanchang, China

ARTICLE INFO

Article history:
Received 22 May 2007
Received in revised form
4 April 2008
Accepted 11 April 2008
Available online 25 April 2008

Keywords:
Empirical distribution function
Permutations
Simulation study
Sudden infant death syndrome
Test of spatial randomness

ABSTRACT

In this article, we extended the empirical distribution function based test statistic I_k of Skaug and Tjostheim [1993. Nonparametric test of serial independence based on the empirical distribution function. Biometrika 80, 591–602] in the time series setting to D_n for spatial lattice data and derived the asymptotic distribution of the proposed test statistic D_n under the null hypothesis of spatial independence. The size and power of the proposed test statistic under conditional autoregressive model (CAR) were simulated. We applied D_n, Moran's I and Geary's c to the transformed and well-studied sudden infant death syndrome data from North Carolina and found that D_n produced a much smaller p-value in testing spatial independence.

1. Introduction

Data collected from studies in many fields including public health may have explicit spatial information. Snow (1994) in the middle of 19th century used spatial patterns to demonstrate the link of cholera and drinking water prior to the knowledge of etiology of the disease. Along with the rapid advance of computer technology, geography information systems (GIS) have been getting popular recently in public health studies for descriptive presentations of data with spatial information. Analytic statistical methods are yet to be implemented in many main stream GIS software packages.

Spatial statistics methods are generally classified into three categories: geostatistics, methods for lattice data and models for point processes (Cressie, 1993). Geostatistics methods were originally developed for mining (Krige, 1951) for data $Z(s)$ observed on spatial location s, where s varies continuously in the space of interest. For example, $Z(s)$ can be the air pollution level at location s. The location may be any possible point in the continuous space of interest. The variable $Z(s)$ itself can be continuous or discrete. For lattice data, the area of interest is a discrete space. For example, $Z(s)$ can be the mortality rate of county s, where s is from a discrete space of finite or countably many. Similar to the random variable in the continuous space, $Z(s)$ may be discrete or continuous (Besag, 1974). Many smoothing techniques were studied in the literature (Kafadar, 1999). Methods for the spatial point processes assume that the location s itself is generated from a stochastic mechanism. For example, the place where a particular disease is observed is a realization of a stochastic process (Zimmerman, 1993).

Identifying and quantifying spatial dependence in terms of spatial patterns and autocorrelation are important in the applications of spatial statistical analysis in public health (Anderson and Titterington, 1997; Kammann and Wand, 2003). In this article, we proposed and studied a test statistic, D_n, which is based on empirical distribution function for lattice data. In our application,

* Corresponding author at: Division of Biostatistics, School of Public Health, University of Texas, 1200 Herman Pressler, Suite 1006, Houston, TX 77030, USA. Tel.: +1 713 500 9270.
E-mail address: dejian.lai@uth.tmc.edu.

0378-3758/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2008.04.019

D_n provided a much smaller p-value than that of Moran's I and Geary's c. Reviews for spatial analysis in epidemiology and public health are available in Marshall (1993), Moore and Carpenter (1999).

2. Test statistic and its asymptotic distribution

The proposed test statistic is based on empirical distribution function, which is reviewed briefly first in this section. Let $\{x_i\}$ be a stochastic process. The empirical distribution function of $\{x_i\}$ is defined as

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leqslant x), \tag{1}$$

where $I(x_i \leqslant x) = 1$ if $x_i \leqslant x$ and $I(x_i \leqslant x) = 0$ otherwise. Similarly, for a two dimensional process $\{x_i = (x_i^{(1)}, x_i^{(2)})\}$, the empirical distribution function is defined as

$$F_n(x, y) = \frac{1}{n}\sum_{i=1}^{n} I(x_i^{(1)} \leqslant x) I(x_i^{(2)} \leqslant y). \tag{2}$$

In time series setting, a lag k empirical distribution function is defined as

$$F_k(x, y) = \frac{1}{n-k}\sum_{i=k+1}^{n} I(x_{i-k} \leqslant x) I(x_i \leqslant y). \tag{3}$$

Let

$$I_k = \frac{1}{n-k}\sum_{i=k+1}^{n} \{F_k(x_{i-1}, x_i) - F_k(x_{i-1}, \infty) F_k(\infty, x_i)\}^2. \tag{4}$$

Skaug and Tjostheim (1993) proposed

$$ST_n = (n-1)\sum_{k=1}^{p} I_k, \tag{5}$$

for testing serial independence in time series analysis, where p is the number of paired lags used. These types of test statistics based on empirical distribution function were also found to have reasonable power of distinguish chaotic time series from random series (Lai and Chen, 2002).

Originally, the test statistic based on empirical distribution function for quantifying dependence among multivariate random variables was introduced by Blum et al. (1961) using the Cramer–von Mises type of distance of distributions (Hoeffding, 1948). In time series context, it was extended by Skaug and Tjostheim (1993, 1996) to testing serial independence, which was then recently generalized by Hong (1998) via various weighting schemes. Further studies on the empirical distribution function for testing serial independence were presented in Delgado (1996) and Delgado and Mora (2000). A comprehensive review of measuring and testing dependence and independence in time series was provided by Tjostheim (1996).

In this article, we extended the test statistic I_k for time series into the lattice case for testing spatial dependence of spatially observed data. For this purpose, we define

$$D_n = \frac{1}{|N|}\sum_{i,j} \{F^*(x_i, x_j) - F^*(x_i, \infty) F^*(\infty, x_j)\}^2 \phi(i, j), \tag{6}$$

where $N = \{(i, j) : i, j = 1, 2, \ldots, n$ and sites i and j are in the same neighborhood$\}$, $|N|$ is the number of distinct pairs in N, $\phi(i, j) = 1$ if sites i and j are in the same neighborhood, $\phi(i, j) = 0$ if sites i and j are not in the same neighborhood, further, let $\phi(i, i) = 0$, $\sum_{i,j}$ denotes the summation of all possible distinct pairs, and

$$F^*(x, y) = \frac{1}{|N|}\sum_{i,j} I(x_i \leqslant x) I(x_j \leqslant y) \phi(i, j).$$

A special case of N in time series is that we consider sites i and j are in the same neighborhood if $|i - j| = 1$. In this case, D_n become I_1. Similarly, D_n can be I_k, $k = 2, \ldots, p$.

The asymptotic distribution of D_n is derived in the Appendix. As we can see from the Appendix, D_n is in fact a weight U-statistic with degenerative kernels. Central limit theorem for U-statistic with some non-degenerative kernels on lattice data was investigated by Sajjan (2000). It is shown in the Appendix, as $n \to \infty$ (so is $|N| \to \infty$)

$$|N| D_n \to \sum_{k,l}^{\infty} 1/(kl\pi^2)^2 w_{k,l}^2 \quad \text{in distribution}, \tag{7}$$