Computational Statistics and Data Analysis 95 (2016) 83–94

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

On Moran's *I* coefficient under heterogeneity

Tonglin Zhang[a,*], Ge Lin[b]

[a] *Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2067, United States*

[b] *Department of Environmental and Occupational Health, University of Nevada, Las Vegas, NV 89154, United States*

ARTICLE INFO

Article history:
Received 25 December 2014
Received in revised form 28 July 2015
Accepted 24 September 2015
Available online 9 October 2015

Keywords:
Clustering or clusters
Heterogeneity
Martingale central limit theorem
Moran's *I* coefficient
Permutation test
Spatial autocorrelation

ABSTRACT

Moran's *I* is the most popular spatial test statistic, but its inability to incorporate heterogeneous populations has been long recognized. This article provides a limiting distribution of the Moran's *I* coefficient which can be applied to heterogeneous populations. The method provides a unified framework of testing for spatial autocorrelation for both homogeneous and heterogeneous populations, thereby resolving a long standing issue for Moran's *I*. For Poisson count data, a variance adjustment method is provided that solely depends on populations at risk. Simulation results are shown to be consistent with theoretical results. The application of Nebraska breast cancer data shows that the variance adjustment method is simple and effective in reducing type I error rates, which in turn will likely reduce potential misallocation of limited resources.

1. Introduction

In the past 20 years, improved GIS and computer technologies have led to a rapid expansion of statistical methods for the analysis of spatial data. The concept and the usage of computational intensive methods, such as Markov random field methods (Green and Richardson, 2002), geostatistical methods (Gneiting, 2002; Kelsall and Wakefield, 2002; Stein, 2005), and Bayesian disease mapping methods (Waller et al., 1997; Wakefield and Morris, 2001) have developed rapidly. A growing number of social and health scientists have taken up the use of the sophisticated technology and new methodologies of spatial analysis in their empirical work (Best et al., 2000; Goodchild et al., 2000; Pickle et al., 2005). Moran's *I* is the most widely used test statistic in spatial statistical literature, and it has been included in major commercial geographic information systems (e.g., **ArcGIS**, **MapInfo**, **Intergraph**, **Imagine**), spatial analysis packages (e.g., **CrimeStat**, **GeoDa**, **TerraSeer**), and some statistical packages (e.g., **MatLab**, **R**, **S+**, **SAS**).

Moran's *I* coefficient (Moran, 1948) is defined as

$$I = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}(X_i - \bar{X})(X_j - \bar{X})}{S_{0m} b_{2m}}, \tag{1}$$

where X_i is the variable of interest in region i $(i = 1, \ldots, m)$, $\bar{X} = \sum_{i=1}^{m} X_i/m$, $S_{0m} = \sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}$, $b_{2m} = \sum_{i=1}^{m}(X_i - \bar{X})^2/m$ and w_{ij} with $w_{ii} = 0$ is an element of the spatial weight matrix. Moran's I mostly ranges between -1 and 1, but it can be outside of $[-1, 1]$ in extreme cases (Arbia, 2014, P. 2). The absolute value of I is bounded by the square root of the ratio between the variance of spatially lagged value and the variance of observed values (Arbia, 1989). It has also been shown

[*] Corresponding author.
E-mail addresses: tlzhang@purdue.edu (T. Zhang), ge.kan@unlv.edu (G. Lin).

http://dx.doi.org/10.1016/j.csda.2015.09.010
0167-9473/© 2015 Elsevier B.V. All rights reserved.

that the value of Moran's I varies from the largest and the least eigenvalues of the weight matrix (Griffith, 1988). These conclusions can provide a way to understand the possible ranges for the values of Moran's I. In applications, a significant positive autocorrelation indicates the existence of either high-value or low-value clustering, while a negative autocorrelation indicates a tendency toward the juxtaposition of high values next to low values. If there is no spatial dependence, the expected value of I is equal to $-1/(m-1)$, which is close to 0 if m is large.

The null distribution of Moran's I is derived from the assumption that distributions of X_i are homogeneous. The p-value of Moran's I is based on its z-score defined by $Z(I) = [I - E(I)]/\sqrt{V(I)}$, where $E(I)$ and $V(I)$ are the mean and variance respectively. There are two ways to define the null hypothesis. The first assumes that $X_1, \ldots, X_m$ are independently and identically distributed (i.i.d.) and the second assumes that $X_1, \ldots, X_m$ are obtained from a random permutation of observed values. The validity of the asymptotic $N(0, 1)$ for $Z(I)$ in the i.i.d. case is evident from Sen (1976), but this assumption may not be valid for variables based on rate when population sizes among area units vary substantially (Besag and Newell, 1991). Although several alternative methods have been proposed (Assuncao and Reis, 1999; Oden, 1995; Waldhor, 1996; Whittemore et al., 1987), for reasons listed below, many still use Moran's I and ignore the potential problems associated with heterogeneous populations. Heterogeneity continues to be the central problem for epidemic models and data (Millison et al., 1994; Zhang and Lin, 2009).

First, all the alternative methods tend to introduce a new test statistic based on regional counts and populations at-risk, and none of them have received wide acceptance. For instance, the population weighted Moran's I proposed by Oden (1995) is essentially a spatially χ^2 statistic, which is not always effective to account for heterogeneous populations (Assuncao and Reis, 1999). Second, some data (e.g., confidential and historical) are only available in rates, for which most alternative methods cannot be used. Third, it is not clear in which situation the population heterogeneity problem will become a serious concern, as such that Moran's I should not be used. Finally, the proliferation of Moran's I in various software packages makes it a candidate for potential misuse as one might simply be unaware of the problem.

In this paper, we attempt to resolve the population heterogeneity issue by providing a unified statistical framework through the limiting distribution of the Moran's I coefficient. Since heterogeneous populations could cause variance inflation, a large sample distribution of Moran's I under heterogeneity should be able to gauge and adjust variance inflation or deflation for $Z(I)$. In the following sections, we first derive a limiting distribution of Moran's I, and then demonstrate its use with two numerical examples in Section 3 and a case study in Section 4. Finally, we offer some concluding remarks.

2. Notation and main result

Notation. Assume a study area partitioned into m regions. Let X_i be the variable of interest from the ith region. Suppose $X_1, \ldots, X_m$ are assumed to be independent. Let $\mu_i = E(X_i)$, $\sigma_i^2 = V(X_i)$, $\kappa_i = E[(X_i - \mu_i)^4]$, $\bar{\mu} = \sum_{i=1}^m \mu_i/m$, $\bar{\sigma}^2 = \sum_{i=1}^m \sigma_i^2/m$. Let $X_{im} = (X_i - \bar{\mu})/\bar{\sigma}$, $\bar{X} = \sum_{i=1}^m X_i/m$, $\bar{X}_{\cdot m} = \sum_{i=1}^m X_{im}/m$, $\mu_{im} = E(X_{im}) = (\mu_i - \bar{\mu})/\bar{\sigma}$, $\sigma_{im}^2 = V(X_{im}) = \sigma_i^2/\bar{\sigma}^2$ and $\kappa_{im} = E[(X_{im} - \nu_i)^4] = \kappa_i/\bar{\sigma}^4$, $i = 1, \ldots, m$. Then $\sum_{i=1}^m \mu_{im} = 0$ and $\sum_{i=1}^m \sigma_{im}^2/m = 1$. For a positive integer k, we write $b_{km} = \sum_{i=1}^m (X_i - \bar{X})^k/m$, $\tilde{b}_{km} = \sum_{i=1}^m (X_{im} - \bar{X}_{\cdot m})^k/m$, $\eta_{km} = \sum_{i=1}^m (\mu_i - \bar{\mu})^k/m$ and $\tilde{\eta}_{km} = \sum_{i=1}^m \mu_{im}^k/m$. Then $\tilde{b}_{km} = b_{km}/\bar{\sigma}^k$ and $\tilde{\eta}_{km} = \eta_{km}/\bar{\sigma}^k$. We write $w_{i\cdot} = \sum_{j=1}^m w_{ij}/m$, $w_{\cdot i} = \sum_{j=1}^m w_{ji}/m$, $S_{0m} = \sum_{i=1}^m \sum_{j=1}^m w_{ij}$, $S_{1m} = \sum_{i=1}^m \sum_{j=1}^m w_{ij}^2$, $S_{2m} = m^2 \sum_{i=1}^m w_{i\cdot}^2$, $\psi_m = S_{0m}/m$ and $\omega_m^2 = 2S_{1m}/m$. We denote $\tilde{\psi}_m = \sum_{i=1}^m \sum_{j=1}^m |w_{ij}|/m$,

$$\theta_m = \frac{\sqrt{m}\sum_{i=1}^m\sum_{j=1}^m w_{ij}(\mu_i - \bar{\mu})(\mu_j - \bar{\mu})}{\sum_{i=1}^m \sigma_i^2} = \frac{1}{\sqrt{m}} \sum_{i=1}^m \sum_{j=1}^m w_{ij}\mu_{im}\mu_{jm} \tag{2}$$

and

$$\tau_m^2 = \frac{2m\sum_{i=1}^m\sum_{j=1}^m w_{ij}^2\sigma_i^2\sigma_j^2}{\left[\sum_{i=1}^m \sigma_i^2\right]^2} + \frac{4m\sum_{k=1}^m \sigma_k^2 \left[\sum_{j=1}^m (w_{kj} + w_{\cdot j})(\mu_j - \bar{\mu})\right]^2}{\left[\sum_{i=1}^m \sigma_i^2\right]^2}$$
$$= \frac{2}{m}\sum_{i=1}^m\sum_{j=1}^m w_{ij}^2\sigma_{im}^2\sigma_{jm}^2 + \frac{4}{m}\sum_{k=1}^m \sigma_{km}^2\left[\sum_{j=1}^m (w_{kj} + w_{\cdot j})\mu_{jm}\right]^2. \tag{3}$$

Throughout the paper, we assume that the fourth moment of X_i exists for all $i \le m$. We write $\xrightarrow{P}$ as convergence in probability and $\xrightarrow{L}$ as convergence in law (or in distribution).

Main Result. We impose the following regularity conditions for our main result:

(C1) For any i and j, $w_{ii} = 0$.
(C2) For any fixed $i \le m$, there is a constant C such that $\sum_{j=1}^m |w_{ij}| \le C$.