Available online at www.sciencedirect.com

ScienceDirect

Statistics and Probability Letters 78 (2008) 1656–1661

STATISTICS & PROBABILITY LETTERS

www.elsevier.com/locate/stapro

Limiting distribution of the G statistics

Tonglin Zhang

Department of Statistics, Purdue University, 250 North University Street, West Lafayette, IN 47907-2066, United States

Received 11 July 2006; received in revised form 16 November 2007; accepted 3 January 2008
Available online 15 January 2008

Abstract

The G statistic and its local version have been used extensively in spatial data analysis. The paper proves the asymptotic normality of the G statistic. Theorems in this paper imply that the regular permutation test for the G statistic is valid.

1. Introduction

Moran's I (1948), Geary's c (1954) and Getis and Ord 's G (1992) are three well-known global test statistics for spatial clustering or autocorrelation. In their initial publications, Moran's I and Geary's c were both conjectured as approximately normally distributed under the null hypothesis of no spatial clustering or autocorrelation when the sample size is large. Sufficient conditions for the asymptotic normality of Moran's I and Geary's c were provided by Sen (1976) about 20 years after they were first published. Although the G statistic was first proposed by Getis and Ord (1992), no proof of asymptotic normality has been given until now. The purpose of this paper is to provide a proof for the asymptotic normality of the G statistic under very weak regularity conditions.

In addition, this article also shows that the normal approximation of the local G_i statistic is inappropriate, especially when there is excessive skewness in observations. Consequently, care should be exercised when using a normal approximation for the distribution of the local G_i statistic as it commonly done in practice.

2. Main result

Let X_i be the variable of interest in region i ($i = 1, \ldots, m$), Getis–Ord's G statistic (Getis and Ord, 1992) is defined as

$$G = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} X_i X_j}{\sum_{i=1}^{m}\sum_{j=1, j\neq i}^{m} X_i X_j} = \frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} X_i X_j}{m^2 b_1^2 - m b_2} \tag{1}$$

where $b_k = \sum_{i=1}^{m} X_i^k / m$ and w_{ij} with $w_{ii} = 0$ is the spatial weight between regions i and j. Even though the spatial weight w_{ij} may not be symmetric, we can assume that they are in this paper since this assumption does not affect

E-mail address: tlzhang@stat.purdue.edu.

0167-7152/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.spl.2008.01.023

the asymptotic properties of the statistic. The spatial weight w_{ij} can be defined by either spatial adjacency or spatial distance.

Suppose that $X_1, \dots, X_m$ are iid nondegenerated random variables with finite fourth moments. Let $\mu = E(X_i)$, $\sigma^2 = V(X_i)$, $\kappa_3 = E(X_i^3)$, $\kappa_4 = E(X_i^4)$. Denote $S_{0m} = \sum_{i=1}^m \sum_{j=1}^m w_{ij}$, $S_{1m} = 2\sum_{i=1}^m \sum_{j=1}^m w_{ij}^2$, $S_{2m} = 4\sum_{i=1}^m w_{i\cdot}^2$, $\tilde{S}_{0m} = \sum_{i=1}^m \sum_{j=1}^m \tilde{w}_{ij}$, $\tilde{S}_{1m} = 2\sum_{i=1}^m \sum_{j=1}^m \tilde{w}_{ij}^2$ and $\tilde{S}_{2m} = 4\sum_{i=1}^m \tilde{w}_{i\cdot}^2$, where $w_{i\cdot} = \sum_{j=1}^m w_{ij}$, $\tilde{w}_{i\cdot} = \sum_{j=1}^m \tilde{w}_{ij}$, $\tilde{w}_{ij} = w_{ij} - S_{0m}/[m(m-1)]$ if $i \neq j$ and $\tilde{w}_{ii} = 0$. Write $\xrightarrow{L}$ as convergence in distribution or in law and $\xrightarrow{P}$ as convergence in probability as $m \to \infty$.

Adopting the standard arguments for the permutation test (Cliff and Ord, 1981, p 14), the mean and variance of the G statistic are obtained by considering the set of $m!$ random permutations of the observed values of $X_1, X_2, \dots, X_m$. Under the null hypothesis of spatial independence, these permutations are equally likely. Denote $E_R(\cdot)$ and $V_R(\cdot)$ as the mean and variance under these random permutations. Getis and Ord obtain the following expressions:

$$E_R(G) = \frac{S_{0m}}{m(m-1)} \tag{2}$$

and

$$\begin{aligned} E_R(G^2) = {} & \frac{S_{1m}(mb_2^2 - b_4)}{(m^2b_1^2 - mb_2)^2(m-1)} + \frac{(S_{2m} - 2S_{1m})(m^2b_1^2b_2 - 2mb_1b_3 - mb_2^2 + 2b_4)}{(m^2b_1^2 - mb_2)^2(m-1)(m-2)} \\ & + \frac{(S_{0m}^2 + S_{1m} - S_{2m})(m^3b_1^4 - 6m^2b_1^2b_2 + 8mb_1b_3 + 3mb_2^2 - 6b_4)}{(m^2b_1^2 - mb_2)^2(m-1)(m-2)(m-3)}. \end{aligned} \tag{3}$$

The permutation variance of the G statistic can be easily obtained by using $V_R(G) = E_R(G^2) - [E_R(G)]^2$.

Let

$$\tilde{G} = \sum_{i=1}^m \sum_{j=1}^m w_{ij} X_i X_j \tag{4}$$

be the numerator of the G statistic given by Eq. (1). Note the denominator $m^2b_1^2 - mb_2$ is invariant under random permutations. Under the random permutation test scheme, the mean and mean square of $\tilde{G}$ can be obtained by

$$E_R(\tilde{G}) = \frac{S_{0m}}{m(m-1)} \sum_{i=1}^m \sum_{j=1, j\neq i}^m X_i X_j \tag{5}$$

and

$$\begin{aligned} E_R(\tilde{G}^2) = {} & \frac{S_{1m}(mb_2^2 - b_4)}{(m-1)} + \frac{(S_{2m} - 2S_{1m})(m^2b_1^2b_2 - 2mb_1b_3 - mb_2^2 + 2b_4)}{(m-1)(m-2)} \\ & + \frac{(S_{0m}^2 + S_{1m} - S_{2m})(m^3b_1^4 - 6m^2b_1^2b_2 + 8mb_1b_3 + 3mb_2^2 - 6b_4)}{(m-1)(m-2)(m-3)}. \end{aligned} \tag{6}$$

Again, the permutation variance of the $\tilde{G}$ can also be easily obtained by using $V_R(\tilde{G}) = E_R(\tilde{G}^2) - [E_R(\tilde{G})]^2$.

To prove that G is asymptotically normal and $[G - E_R(G)]/\sqrt{V_R(G)}$ is asymptotically $N(0, 1)$ as $m \to \infty$, I impose the following regularity conditions:

(C1) $w_{ij} = w_{ji}$ and $w_{ii} = 0$ for all $i, j \leq m$.
(C2) There is a constant C such that $\sum_{j=1}^m |w_{ij}| \leq C$ for all $i \leq m$.

The regularity conditions (C1) and (C2) are exactly imposed by Sen (1976) when he proves the asymptotic normality of Moran's I and Geary's c statistics. Condition (C1) implies that S_{0m}/m, S_{1m}/m and S_{2m}/m are uniformly bounded. To derive the asymptotic mean and variance of the G statistic under the random permutation test scheme, it is necessary to impose a stronger version of the conditions, which states that the following three limits exist and are all positive: $\gamma_0 = \lim_{m\to\infty} S_{0m}/m$, $\gamma_1^2 = \lim_{m\to\infty} S_{1m}/m = \gamma_1^2$ and $\gamma_2^2 = \lim_{m\to\infty} S_{2m}/m$.

The proof of the asymptotic normality of G statistic under the random permutation test scheme is an application of the Martingale Central Limit Theorem (Billingsley, 1995, P 476). As the denominator is invariant under random permutation test scheme, the asymptotic normality is derived by the proof of the asymptotic normality of the numerator $\tilde{G}$ statistic first and then the implement of the limits of remaining terms in the sense of convergence in probability second.

Let

$$T_m = \frac{1}{\sqrt{m}}[\tilde{G} - E_R(\tilde{G})] = \frac{1}{\sqrt{m}} \sum_{i=1}^{m}\sum_{j=1}^{m} \tilde{w}_{ij}(X_i - \mu)(X_j - \mu) + \frac{2\mu}{\sqrt{m}} \sum_{i=1}^{m} \tilde{w}_{i\cdot}(X_i - \mu).$$

Then $E(T_m) = 0$ and

$$V(T_m) = \frac{\sigma^4 \tilde{S}_{1m}}{m} + \frac{\mu^2\sigma^2\tilde{S}_{2m}}{m} \to \sigma^4\gamma_1^2 + \mu^2\sigma^2(\gamma_2^2 - 4\gamma_0^2)$$

as $m \to \infty$. Therefore, we have the following theorem.

Theorem 1. $T_m \xrightarrow{L} N(0, \sigma^4\gamma_1^2 + \mu^2\sigma^2(\gamma_2^2 - 4\gamma_0^2))$.

Proof. The proof is just an application of the Martingale Central Limit Theorem. Suppose that $\{Y_{ij} : j = 1, 2, \ldots, \}$ is a martingale with respect to σ-fields $\{\mathcal{F}_{ij} : j = 1, 2, \ldots, \}$, let $Z_{ij} = Y_{ij} - Y_{i,j-1}$ be the martingale difference and let $\omega_{ij}^2 = E(Z_{ij}^2|\mathcal{F}_{i,j-1})$. The Martingale Central Theorem says that if $\sum_{j=1}^{\infty}\omega_{ij}^2 \xrightarrow{P} \omega^2 > 0$ and also the Lindeberg–Feller condition that $\sum_{j=1}^{\infty} E(Z_{ij}^2 I_{|Z_{ij}|\geq\epsilon}) \to 0$ for every $\epsilon > 0$ as $i \to \infty$, then $\sum_{j=1}^{\infty} Z_{ij} \xrightarrow{L} N(0, \omega^2)$ as $i \to \infty$.

Let

$$T_{mk} = \frac{1}{\sqrt{m}} \sum_{i=1}^{k}\sum_{j=1}^{k} \tilde{w}_{ij}(X_i - \mu)(X_j - \mu) + \frac{2\mu}{\sqrt{m}} \sum_{i=1}^{k} \tilde{w}_{i\cdot}(X_i - \mu)$$

for $k = 1, \ldots, m$ and let $T_{m0} = 0$. Let $\mathcal{F}_{mk} = \sigma(T_{m0}, T_{m1}, \ldots, T_{mk})$ be the σ-algebra generated by $T_{m0}, \ldots, T_{mk}$ for $k = 0, 1, \ldots, m$. Then $\{T_{mk} : k = 0, 1, \ldots, m\}$ is a martingale with respect to $\{\mathcal{F}_{mk} : k = 0, 1, \ldots, m\}$. Let $Z_{mk} = T_{mk} - T_{m,k-1}$ be the martingale difference. Define

$$\begin{aligned}\omega_{mk}^2 &= E(Z_{mk}^2|\mathcal{F}_{m,k-1}) \\ &= \frac{4\sigma^2}{m}\sum_{i=1}^{k-1}\tilde{w}_{ik}^2(X_i - \mu)^2 + \frac{4\mu^2\sigma^2\tilde{w}_{k\cdot}^2}{m} \\ &\quad + \frac{8\sigma^2\mu\tilde{w}_{k\cdot}}{m}\sum_{i=1}^{k-1}\tilde{w}_{ik}(X_i - \mu) + \frac{4\sigma^2}{m}\sum_{i=1}^{m}\sum_{j=1, j\neq i}^{m}\tilde{w}_{ik}\tilde{w}_{jk}(X_i - \mu)(X_j - \mu).\end{aligned}$$

Let $\omega_m^2 = \sum_{k=1}^{m}\omega_{mk}^2$. Then, we have

$$\omega_m^2 = \frac{4\sigma^2}{m}\sum_{k=1}^{m}\sum_{i=1}^{k-1}\tilde{w}_{ik}^2(X_i - \mu)^2 + \frac{\mu^2\sigma^2\tilde{S}_{2m}}{m} + R_{1m} + R_{2m}, \tag{7}$$

where

$$R_{1m} = \frac{4\sigma^2\mu}{m}\sum_{k=1}^{m}\sum_{i=1}^{m}\tilde{w}_{k\cdot}\tilde{w}_{ik}(X_i - \mu)$$

and

$$R_{2m} = \frac{4\sigma^2}{m}\sum_{k=1}^{m}\sum_{i=1}^{k-1}\sum_{j=1,j\neq i}^{k-1}\tilde{w}_{ik}\tilde{w}_{jk}(X_i-\mu)(X_j-\mu)$$
$$= \frac{4\sigma^2}{m}\sum_{i=1}^{m-1}\sum_{j=1,j\neq i}^{m-1}(X_i-\mu)(X_j-\mu)\left[\sum_{k=\max(i,j)+1}^{m}\tilde{w}_{ik}\tilde{w}_{jk}\right].$$

Since $E(R_{1m})=0$ and

$$E(R_{1m}^2) = \frac{16\sigma^6\mu^2}{m^2}\sum_{i=1}^{m}\left[\sum_{k=1}^{m}\tilde{w}_{k\cdot}\tilde{w}_{ik}\right]^2 \leq \frac{16\sigma^6\mu^2C^4}{m},$$

we have $R_{1m}\xrightarrow{P}0$ as $m\to\infty$ by the Chebyshev inequality (Billingsley, 1995, P 80). Similarly, we can also prove that $R_{2m}\xrightarrow{P}0$ by the Chebyshev inequality. Since $\tilde{S}_{1m}/m = S_{1m}/m + o(1)$ and $\tilde{S}_{2m} = S_{2m}/m - 4S_{0m}/m^2$, we can prove that the first term in (7) goes to $\sigma^4\gamma_1^2$ and the second term in (7) goes to $\mu^2\sigma^2(\gamma_2^2-4\gamma_0^2)$ in probability by the Chebyshev inequality. Thus, $\omega_m^2 \xrightarrow{P} \sigma^4\gamma_1^2+\mu^2\sigma^2(\gamma_2^2-4\gamma_0^2)$.

The rest of the proof is to check the Lindeberg–Feller condition, which can be implied by the Lyapounov Condition (Billingsley, 1995, P 359–362). Particularly in this case, the Lyapounov Condition says that there is an $\epsilon>0$ such that $\lim_{m\to\infty}\sum_{i=1}^{m}E(Z_{mk}^{2+\epsilon})=0$. This can be confirmed by taking $\epsilon=2$. In this case, we have

$$\sum_{i=1}^{m}E(Z_{mk}^4) = E\left\{\sum_{k=1}^{m}\frac{16(X_k-\mu)^4}{m^2}\left[\sum_{i=1}^{k-1}\tilde{w}_{ij}(X_i-\mu)+\mu\tilde{w}_{k\cdot}\right]^4\right\}$$
$$\leq \frac{128\kappa_4}{m^2}\sum_{k=1}^{m}\left\{E\left[\sum_{i=1}^{k-1}\tilde{w}_{ij}(X_i-\mu)\right]^4+\mu^4\tilde{w}_{k\cdot}^4\right\}$$
$$\leq \frac{128C^4\kappa_4(\kappa_4+\sigma^4+\mu^4)}{m}.$$

Therefore by the Martingale Central Limit Theorem, $T_m\xrightarrow{L}N(0,\sigma^4\gamma_1^2+\mu^2\sigma^2(\gamma_2^2-4\gamma_0^2))$. ◇

According to Getis and Ord (1992), G statistic is recommended to test for spatial clustering or autocorrelation for nonnegative random variables, the following theorem states the asymptotic normality including this case.

Corollary 1. *If $\mu>0$, then $m^{3/2}(G-S_{0m}/m^2)\xrightarrow{L}N(0,\sigma^4\gamma_1^2/\mu^2+\sigma^2(\gamma_2^2-4\gamma_0^2))$ as $m\to\infty$.*

Proof. This is obvious since

$$m^{3/2}\left(G-\frac{S_{0m}}{m^2}\right) = \frac{1}{b_1^2-b_2/m}T_m + \frac{S_{0m}}{(m-1)\sqrt{m}},$$

$S_{0m}/[(m-1)\sqrt{m}] = O(m^{-1/2})$, $b_1\xrightarrow{P}\mu>0$ and $b_2/m\xrightarrow{P}0$. ◇

Corollary 2. $[G-E_R(G)]/\sqrt{V_R(G)}\xrightarrow{L}N(0,1)$ *as $m\to\infty$.*

Proof. It is clear that $b_1\xrightarrow{P}\mu$, $b_2\xrightarrow{P}\mu^2+\sigma^2$, $b_3\xrightarrow{P}\kappa_3$ $b_4\xrightarrow{P}\kappa_4$, and

$$\frac{G-E_R(G)}{\sqrt{V_R(G)}} = \frac{\tilde{G}-E_R(\tilde{G})}{\sqrt{V_R(\tilde{G})}}.$$

By ignoring the small order terms, we have

$$V_R\left(\frac{\tilde{G}}{\sqrt{m}}\right) = \frac{S_{1m}(b_2-b_1^2)^2}{m} + \frac{S_{2m}(b_1^2b_2-b_1^4)}{m} + \frac{4S_{0m}^2(b_1^4-b_1^2b_2)}{m^2} + o_p(1).$$

This implies

$$V_R\left(\frac{\tilde{G}}{\sqrt{m}}\right) \xrightarrow{P} \sigma^4\gamma_1^2 + \mu^2\sigma^2(\gamma_2^2 - 4\gamma_0^2)$$

and $[\tilde{G} - E_R(\tilde{G})]/\sqrt{V_R(\tilde{G})} \xrightarrow{L} N(0, 1)$. ◇

Remark. The existence of the fourth moment of X_i is the minimum requirement since $V_R(\tilde{G}/\sqrt{m})$ has a term of b_4.

3. The local statistic

For a reference location i, local Getis-Ord's G_i statistic in its standardized form (Getis and Ord, 1992) is

$$G_i = \frac{\sum_{j=1}^{m} w_{ij} X_j}{\sum_{j=1, j\neq i}^{m} X_i}$$

or

$$G_i^* = \frac{\sum_{j=1}^{m} w_{ij} X_j}{\sum_{j=1}^{m} X_i},$$

where w_{ij} with $w_{ii} = 0$ is also the spatial weight between regions i and j as before. The forms of G_i and G_i^* require that the underlying variable X_i be nonnegative.

The asymptotic normality assumption of the local G_i or G_i^* statistic has been used extensively for computing their p-values. In the original paper by Getis and Ord (1992), they suggest to compute the p-values of the G_i statistic by a z-test. In their method, they assume that G_i is approximately normally distributed with both the expected value and the variance being computed under random permutations. The following example displays a case when the asymptotic normality of G_i statistic is invalid.

Example. Let $A \subseteq R^2$ be the set of grid points with both the absolute values of the horizontal and vertical coordinates less than or equal to k. Then A is a $(2k + 1) \times (2k + 1)$ lattice centered at $(0, 0)$. We call location i and j neighbors if they are on the same row but next columns or vice versa. Then, points $(-1, 0)$, $(0, -1)$, $(1, 0)$ and $(0, 1)$ are the only four neighbors of point $(0, 0)$. We consider the common rook weight in spatial analysis as $w_{ij} = 1$ if location i and location j are neighbors and $w_{ij} = 0$ otherwise. Let $X_{(a,b)}$ be the random variable generating at point (a, b). If location i indicates $(0, 0)$, then

$$G_i = \frac{X_{(-1,0)} + X_{(0,-1)} + X_{(0,1)} + X_{(1,0)}}{m\bar{X} - X_{(0,0)}},$$

where $m = (2k + 1)^2$ is the total number of lattice points. If $X_{(a,b)}$ are iid Gamma(α, β) random variables, then $\beta G_i/(m\alpha)$ weakly converges to Gamma$(4\alpha, \beta)$ as $k \to \infty$, which indicates that the normality assumption is invalid. The extreme case happens when α is very close to 0, in which the underlying distribution is very skewed. The p-value obtained from the asymptotic normality assumption is significantly different from its true value. Similar conclusion can also be drawn for the local G_i^* statistic.

4. Concluding remarks

Based on the Martingale Central Limit Theorem, this paper shows that the sufficient conditions for the asymptotic normality of the G statistic are almost identical to those of the I and c statistics provided by Sen (1976). Both the numerators of the G statistic and the I statistic are in quadratic forms that include only the cross product terms. The

I statistic is centered by its sample mean while G is not. Even though both are able to test for spatially dependence structures, the I statistic focuses on discriminating spatial clustering and negative autocorrelation but the G statistic focuses on discriminating high value clustering from other spatial structures. Like local I_i statistic, the distribution of the local G_i is not approximately normal if the number of adjacent sites is small. The normal approximation of the null distribution is valid if the underlying distributions of X_i is normally distributed (Leung et al., 2003). In real applications, for the similar issues like the local I_i and c_i (Tiefelsdorf, 2002), one needs to be careful about using the normality assumption for the p-value computation the underlying distribution of X_i is serious skewed.

Acknowledgment

It is a pleasure to thank Dr. Ge Lin for numerous conversations I had with him on the background questions of the paper. This paper would have never been written without his help.

References

Billingsley, P., 1995. Probability and Measure. Wiley, New York.

Cliff, A.D., Ord, J.K., 1981. Spatial Processes: Models And Applications. Pion, London.

Geary, R.C., 1954. The contiguity ratio and statistical mapping. The Incorporated Statistician 5, 115–145.

Getis, A., Ord, J., 1992. The analysis of spatial association by use of distance statistics. Geographical Analysis 24, 189–206.

Leung, Y., Mei, C., Zhang, W., 2003. Statistical test for local patterns of spatial association. Environment and Planning A 35, 725–744.

Moran, P.A.P., 1948. The interpretation of statistical maps. Journal of the Royal Statistic Society Series B 10, 243–251.

Sen, A., 1976. Large sample-size distribution of statistics used in testing for spatial correlation. Geographical Analysis 9, 175–184.

Tiefelsdorf, M., 2002. The saddlepoint approximation of Moran's I's and local Moran's I_i's reference distribution and their numerical evaluation. Geographical Analysis 34, 187–206.