Journal of Geographical Systems (2019) 21:237–269
https://doi.org/10.1007/s10109-019-00293-3

ORIGINAL ARTICLE

Spatial autocorrelation for massive spatial data: verification of efficiency and statistical power asymptotics

Qing Luo[1,3] · Daniel A. Griffith[2] · Huayi Wu[1,3]

Received: 24 February 2018 / Accepted: 5 March 2019 / Published online: 25 March 2019

Abstract

Being a hot topic in recent years, many studies have been conducted with spatial data containing massive numbers of observations. Because initial developments for classical spatial autocorrelation statistics are based on rather small sample sizes, in the context of massive spatial datasets, this paper presents extensions to efficiency and statistical power comparisons between the Moran coefficient and the Geary ratio for different variable distribution assumptions and selected geographic neighborhood definitions. The question addressed asks whether or not earlier results for small n extend to large and massively large n, especially for non-normal variables; implications established are relevant to big spatial data. To achieve these comparisons, this paper summarizes proofs of limiting variances, also called asymptotic variances, to do the efficiency analysis, and derives the relationship function between the two statistics to compare their statistical power at the same scale. Visualization of this statistical power analysis employs an alternative technique that already appears in the literature, furnishing additional understanding and clarity about these spatial autocorrelation statistics. Results include: the Moran coefficient is more efficient than the Geary ratio for most surface partitionings, because this index has a relatively smaller asymptotic as well as exact variance, and the superior power of the Moran coefficient vis-à-vis the Geary ratio for positive spatial autocorrelation depends upon the type of geographic configuration, with this power approaching one as sample sizes become increasingly large. Because spatial analysts usually calculate these two statistics for interval/ration data, this paper also includes comments about the join count statistics used for nominal data.

Keywords Moran coefficient · Geary ratio · Efficiency · Power · Geographic configuration · Join count statistics

JEL Classification C12 · C46 · C55

✉ Huayi Wu
wuhuayi@whu.edu.cn

Extended author information available on the last page of the article

1 Introduction

Because georeferenced data, some of which are real time, containing locational information have been continuously returned by a variety of sensors (e.g., public transport vehicles equipped with a global position system, remote sensing satellites, and smartphones) and obtained from more and more open sources (such as government health statistical data and demographic data), the amount of spatial data is increasing at an explosive rate. This kind of data is called big spatial data or big geospatial data. Compared with traditional spatial data, these data have a much bigger volume, much more variety, and much higher velocity, and the tools for processing and analyzing them are more complex (van Zyl 2014; Lee and Kang 2015; Li et al. 2016, Haynes et al. 2018). One feature of this kind of data is massive sample sizes, which is accompanied by the basic problem of still being able to detect the latent spatial autocorrelation (SA) across different geographical structures, and, furthermore, whether or not statistical properties for small-to-medium size datasets pertain to large-to-massive datasets.

For a traditional analysis, two typical statistics that have been devised to quantify the nature and degree of SA are the Moran coefficient (MC) and the Geary ratio (GR), which employ different metrics according to their mathematical expressions. The MC contains a cross-product term in its numerator pertaining to deviations from the mean [see Eq. (1)]; this construction is similar to Pearson's product moment correlation coefficient, *r*, whose spatial counterpart may be a SA parameter of a spatial autoregressive model, which is widely employed by researchers across a range of disciplines. In other words, the MC corresponds to the spatial autoregressive perspective. The GR contains a paired comparison term in its numerator, one in which differences are between observation attribute values [see Eq. (2)]; this quantity is similar to that used to construct a semivariogram in which the geographic variation between two locations is expressed as a difference between two observation attribute values. In other words, the GR corresponds to the geostatistical semivariogram perspective (Legendre and Fortin 1989). Spatial autoregression works with the inverse covariance matrix, whereas geostatistics works directly with the spatial covariant matrix.

Therefore, the comparison is not only for two single statistics, but for two different conceptualizations. Although these two indices were introduced many decades ago by Moran (1950) and Geary (1954), respectively, they were not widely employed in terms of SA indexes until Cliff and Ord (1973, 1981) published their fundamental and pioneering works, in which these two statistics' distributional properties, including their asymptotic normal sampling distributions and power (i.e., the probability of rejecting the null hypothesis when it is not true) comparisons for positive SA of small sample sizes were established in detail. Thereafter, various researches related to these SA statistics began to appear. For example, Griffith (1987, p. 44) first pointed out the relationship function expressing the MC in terms of the GR. Tiefelsdorf and Boots (1995) derived the exact distribution of the MC for small samples, which is a seminal work that helped to establish the novel Moran eigenvector spatial filtering spatial statistics methodology

(Griffith 1996). Anselin (1995, 1996) introduced local indicators of spatial association (LISA) and the Moran scatter plot, which visualizes SA with a regression trend line superimposed on those geographical attribute points appearing in the numerator of the MC distributed across the four quadrants of the plane. Boots (2003) also furnished local SA indices for categorical data. Lee (2001) developed a bivariate spatial correlation coefficient as well as its local form by integrating Pearson's r and the MC. Boots and Tiefelsdorf (2000) investigated the behavior of SA test statistics in three regular tessellations; Bivand et al. (2009) implemented the saddlepoint approximation instead of the normal approximation, and the exact distribution of the MC in the R *spdep* package, which makes power analysis easier because many geographic information system (GIS) software packages do not have this function. Chun (2008), Cheng et al. (2012), Bavaud (2013), and de la Mata and Llano (2013) discussed issues relating to network spatial autocorrelation. More recently, Carrijo and da Silva (2017) devised a modified MC to solve the problem of underestimating real SA when sample size is small; Anselin (2018) extended the Local GR to a multivariate context. Except for those theoretical studies on the statistical properties of these two statistics, the MC and GR often are used as tools in explanatory works for descriptive and visualization purposes. In addition, the MC is used as a tool for the diagnosis of SA in regression modeling (Cliff and Ord 1969, 1970).

For a massive spatial data analysis, the mathematical or statistical properties of the MC and the GR need to be extended to much larger sample sizes on the basis of Cliff and Ord's (1973, 1981) pioneering works. One question asks why a researcher still uses SA coefficients to describe large sample size datasets. Being similar to those summary statistics (e.g., the mean, variance, and median) that portray data from different angles, and that are computed as initial descriptions when a researcher obtains his or her dataset, an SA coefficient can be seen as a summary statistic as well in spatial statistics. Thus, regardless of the sample size, knowing the degree of SA is useful so that researchers can have a first impression of the spatial data at hand. Moreover, calculating this statistic is not the target in a spatial data analysis experiment. Rather, it is a tool for determining subsequent treatments, such as the selection of a spatial model for describing data when a MC value indicates strong positive SA. Otherwise, the selection may be a nonspatial model if the MC is not calculated, or ignored. Consequently, extension of small sample size results to large or massive sample sizes is necessary and is the major purpose of this paper. Specifically, we derive the mathematical proofs of the asymptotic variances of the MC and the GR for different types of random variables, through which the MC is shown to be more efficient than the GR for large sample sizes. We also develop an analytical approach to compare the statistical power of the two statistics for any size dataset.

This article substantiates the findings in Luo et al. (2017), with detailed mathematical derivations and interpretations. It includes a methodology part in Sect. 2, followed by a mathematics section. Section 4 analyzes efficiency in terms of different surface partitionings and distribution conditions. Section 5 compares statistical power. Section 6 discusses relationships between the MC, the GR, and the join count statistics based on the work of Cliff and Ord (1973). This paper also provides results in Sect. 7 for two massive spatial dataset examples to verify the findings of the previous sections. Finally, this paper states conclusions and presents discussions in its last section. Its contributions beyond the 2017 paper are the following: detailed proofs for theorems, an alternative visualization of statistical power, a comment on the join count statistics that are applicable to nominal data, and two empirical examples to validate results.

2 Methodology

Most spatial analysts deal with only a few of the many possible types of random variables (e.g., normal, binomial, Poisson). Furthermore, the geospatial literature suggests that a particular set of geographic configurations furnishes useful insights into, and understanding of, many spatial statistics concepts. These are the topics of this section.

2.1 Distributional assumptions and geographic configurations

Throughout this paper, the following two aspects of postulates are set: one pertains to the types of random variable (i.e., distributional assumptions) and the other pertains to geographic configuration, or surface partitioning. These foci are inspired by Cliff and Ord's (1973, 1981) work, in which the moments of the MC and the GR are derived under normality and randomization assumptions, and power curves have been drawn for several geographical configurations (i.e., circular, rook, queen, queen on torus, and an empirical surface partitioning).

This paper analyzes not only the normal distribution, but also three other specific distributions (i.e., uniform, beta with equal scale parameters less than one, and exponential). It also includes six additional geographical configurations (i.e., linear, hexagonal, maximum planar, the two versions of maximum hexagonal, and rook on torus). Essentially, different distributions render different kurtosis terms, and different geographical configurations produce different connectivity matrices. The following subsections describe details about these cases.

2.1.1 Four types of random variables

The four selected distributions are for continuous random variables; those for discrete random variables are not discussed in this paper. Figure 1 portrays their probability density function plots with their respective kurtosis terms (b_2).

These distributions are selected because they furnish a representative sample of the full range of probability distributions. Specifically, the normal family is the most typical case. Each of the other three distributions has no direct connection with the normal distribution, although the exponential distribution can be subjected to a Box–Cox power transformation that approximates a normal distribution, and none of

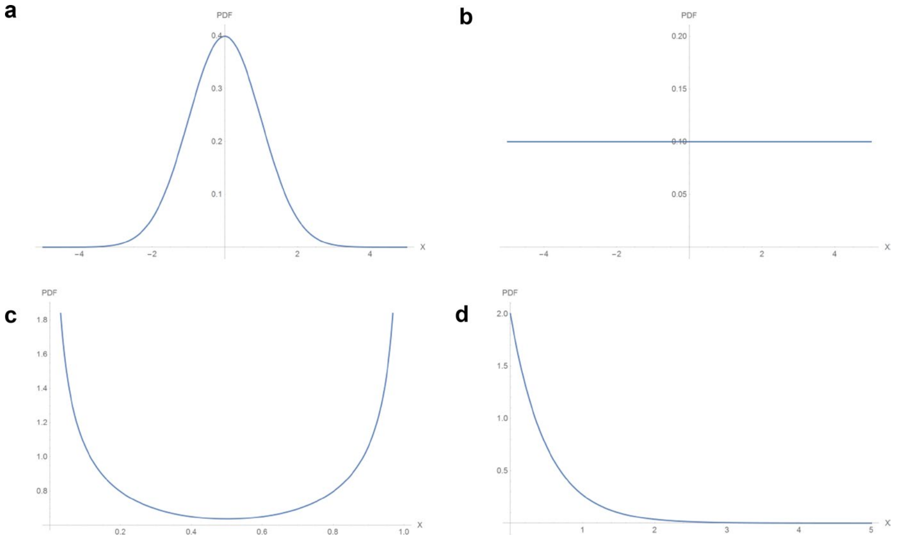

Fig. 1 Probability density function (PDF) plots. **a** Normal distribution, $b_2 = 3$. **b** Uniform distribution, $b_2 = 9/5$. **c** Beta distribution ($\alpha = \beta = 0.5$), $b_2 = 3/2$. **d** Exponential distribution, $b_2 = 9$

these three non-normal distributions has a connection with either of the other two as well.[1] More specifically, the normal distribution represents a non-skewed and proper kurtosis distribution, the uniform distribution depicts a flat distribution within a finite interval, the beta distribution with $\alpha = \beta = 0.5$ (this is the case employed throughout this paper) represents a sinusoidal distribution within a confined interval, and the exponential distribution depicts a skewed and leptokurtic distribution.

2.1.2 Geographic configurations

Ten geographic partitionings are employed: three of them, namely the maximum planar connectivity case, and the two maximum hexagonal cases (with odd and even columns), are theoretically constructed. These settings furnish a relatively comprehensive representation of possible realistic and theoretical configurations. For example, a square rook and a square queen articulation are common in the surface partitioning for remotely sensed images, whereas a hexagonal partitioning often is employed in spatial sampling designs (e.g., Chun and Griffith 2013, pp. 24–29). The linear [each of the internal areal units has two geographic neighbors, while each of the two end areals units has only one neighbor], the circle [a two-dimensional (2-D) counterpart to the linear case], and the torus [a 3-D counterpart to the square rook or queen case] do not relate to empirical landscapes; these 2-D and 3-D cases are configurations in which each areal unit has the same number of neighbors (e.g., for

[1] Refer to *Univariate distribution relationships*: http://www.math.wm.edu/~leemis/chart/UDR/UDR.html.

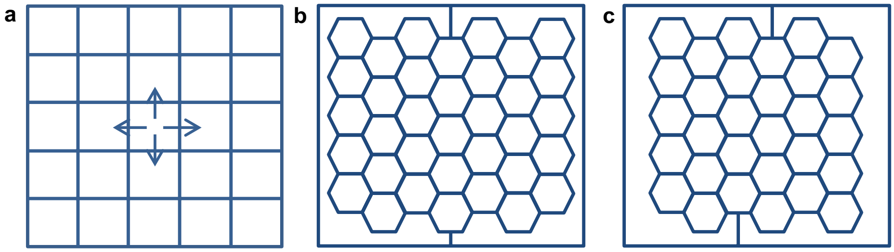

Fig. 2 Selected surface partitionings. **a** A regular square rook configuration. **b** A maximum hexagonal configuration with an odd Q. **c** A maximum hexagonal configuration with an even Q

Table 1 Neighbor sums of selected geographic configurations

Neighbor sum	Regular square rook adjacency ($n = P \times Q$)	Maximum hexagonal partitioning with an odd Q ($n = P \times Q + 2$)	Maximum hexagonal partitioning with an even Q ($n = P \times Q + 2$)
$\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}$	$2(2PQ - P - Q)$	$6PQ$	$6PQ$
$\sum_{i=1}^{n} \left(\sum_{j=1}^{n} c_{ij} \right)^2$	$2(8PQ - 7P - 7Q + 4)$	$2\left(P^2 + Q^2 + 20PQ - 11P - 10Q + 6\right)$	$2\left(P^2 + Q^2 + 20PQ - 11P - 10Q + 8\right)$

a circle, every areal unit has two neighbors, whereas for a torus with rook adjencacy, each cell has four neighbors, and for a torus with queen adjacency, each has eight). And the maximum planar case (Tait and Tobin 2017) can be seen as one possible realization of a planar graph that has the maximum number of edges $3(n-2)$, where n (i.e., the number of areal units) is the number of nodes in a graph. Furthermore, to gain a better understanding of this partitioning, maximum hexagonal cases with different numbers of columns have been designed (the internal linear units of a maximum planar case are replaced with hexagonal cells). Figure 2 portrays three of these situations, and Table 1 lists their corresponding neighbor sums, where P and Q are the number of rows and columns, respectively, in a configuration, $n = P \times Q$ is the number of areal units under study, and $\boldsymbol{C} = \left(c_{ij}\right)_{n \times n}$ is the connectivity matrix, where $c_{ij} = 1$ if areal units i and j are adjacent (i.e., they have a common edge or point, and hence are neighbors), and 0 otherwise; matrix $\boldsymbol{C}$ is symmetric.

In the hexagonal cases, the number of areal units n is no longer $P \times Q$, but rather $P \times Q + 2$, where the additional two areal units are those surrounding the outside of the geographic landscape. These connections between internal hexagons and the outer two areal units are designed to attain the maximum neighbor sums.

To illustrate the variation in different geographic connections and sample sizes, Table 2 presents the extreme eigenvalues of matrix $\boldsymbol{C}$ for each configuration, as well as their corresponding extreme MC and GR values. Discussion of the relationship between these matrix $\boldsymbol{C}$ eigenvalues, λ, and the MC as well as the GR appears in subSect. 5.2.

The analyses for Table 2 employed essentially two different sample sizes, $n \approx 100$ and 10,000. For L and CN-C, $n = 1 \times 100$ or $1 \times 10,000$; for SR, SQ, H, CN-TR, and CN-TQ, $n = 10 \times 10$ or 100×100; for MPC, $n = 1 \times 100 + 2$ or $1 \times 10,000 + 2$; for MH-O and MH-E, $n = 10 \times 11 + 2$ or $100 \times 101 + 2$, and $n = 10 \times 10 + 2$ or

Table 2 Selected eigenvalues and affiliated MC and GR values for different sample sizes and connectivity

n	Values	L	SR	SQ	H	MPC[a]	MH-O[a]	MH-E	CN-C	CN-TR	CN-TQ
100	$\lambda_{\max}$	1.99903	3.83797	7.52048	5.71148	15.6413	6.88544	6.81648	2.00000	4.00000	8.00000
	$\lambda_{\min}$	− 1.99903	− 3.83797	− 3.68251	− 2.86451	− 12.6614	− 4.30963	− 4.18656	− 2.00000	− 4.00000	− 4.00000
102	$\mathrm{MC}_{\max}$	1.00815	1.00042	0.99857	1.02306	0.3393	1.02374	1.00982	0.99803	0.90451	0.85676
	$\mathrm{MC}_{\min}$	− 1.00961	− 1.06610	− 0.53838	− 0.54876	− 0.4934	− 0.73133	− 0.71171	− 1.00000	− 1.00000	− 0.50000
112	$\mathrm{GR}_{\max}$	1.99950	2.13957	1.66595	1.66572	17.1700	3.11885	3.00892	1.98000	1.98000	1.48500
	$\mathrm{GR}_{\min}$	0.00186	0.07242	0.11107	0.08721	0.3373	0.09712	0.10160	0.00195	0.09454	0.14180
10000	$\lambda_{\max}$	2.00000	3.99807	7.99420	5.99661	142.922	15.8901	15.85452	2.00000	4.00000	8.00000
	$\lambda_{\min}$	− 2.00000	− 3.99807	− 3.99613	− 2.99831	− 139.922	− 13.7855	− 13.75050	− 2.00000	− 4.00000	− 4.00000
10002	$\mathrm{MC}_{\max}$	1.00010	1.00888	1.01334	1.01217	0.333	2.5971	2.59104	1.00000	0.99901	0.99852
	$\mathrm{MC}_{\min}$	− 1.00010	− 1.00961	− 0.50710	− 0.50645	− 0.500	− 2.2980	− 2.29220	− 1.00000	− 1.00000	− 0.50000
10102	$\mathrm{GR}_{\max}$	2.00000	2.01949	1.52209	1.51976	1667.167	20.8639	20.77053	1.99980	1.99980	1.49985
	$\mathrm{GR}_{\min}$	0.00000	0.00117	0.00177	0.00127	0.333	0.0013	0.00128	0.00000	0.00099	0.00148

The headers for the configuration columns denote: linear (L), square rook (SR), square queen (SQ), hexagonal (H), maximum planar connectivity (MPC), maximum hexagonal partitioning with an odd number (MH-O) and an even number (MH-E) of columns, and the constant number of neighbors for a circle case (CN-C), a torus rook case (CN-TR), and a torus queen case (CN-TQ)

[a] Not all values in these two columns have five digits after decimal point because of alignment and space limitations

$100 \times 100 + 2$, respectively. Except for MPC, MH-O, and MH-E, all calculated SA index values are well behaved, in that MC + GR ≈ 1, and the maximum eigenvalue is within the interval of the minimum row sum and the maximum row sum of matrix $\boldsymbol{C}$. For example, for the linear case with a smaller sample size, because $n-2$ of the row sums of matrix $\boldsymbol{C}$ are two, and only the first and last rows have a sum of one, the maximum eigenvalue is close to, but slightly less than, two, and also because of the symmetry of those eigenvalues[2] and the zero trace,[3] the minimum eigenvalue is minus the maximum. Furthermore, the summation of the strongest positive SA is $\mathrm{MC}_{\max} + \mathrm{GR}_{\min} = 1.00815 + 0.00186 = 1.01001$, and its negative counterpart is $\mathrm{MC}_{\min} + \mathrm{GR}_{\max} = -1.00961 + 1.99950 = 0.98989$. However, this appealing property no longer holds when the distribution of ones is highly skewed toward two rows that represent connectivities of the outer two units (i.e., MPC, MH-O, and MH-E). This skewness is more serious in the maximum planar case because the peripheral two cells are not only connected with each other, but also with all inner cells; in other words, each of these two rows contains $n-1$ ones, because they are adjacent to every cell except themselves. Hence, a severe unevenly structured adjacency matrix yields considerably large extreme eigenvalues, and MC and GR values; but these undesirable values appear only for the first and last few eigenvalues. Table 6 in Appendix 1 reports the first ten positive and last ten negative λ, and MC and GR values for these three anomalous configurations. It reveals a big difference between selected extreme values (marked with red) and other values in the same row; this difference seems most conspicuous for the MPC case, especially with 10,002 samples. Reasons for these discrepancies include: (1) the number of ones in matrix $\boldsymbol{C}$ for at least one areal unit increases as n increases, and (2) the diameter of the affiliated graph is relatively small.

2.2 Efficiency and variance

There are two "efficiency" measures in statistics. One is used as a criterion that qualifies an estimator—for two unbiased estimators; the one with the smaller variance is more efficient. The other is used in hypothesis testing—when comparing two test procedures; the one that needs fewer observations for a given power is more efficient. This paper uses the first "efficiency" measure, but for another purpose, namely comparing two statistics rather than estimators.

An additional reason for pursuing this efficiency comparison is that Cliff and Ord (1969, p. 45) point out that the variance of the MC is "less affected by the distribution of the sample data" than the variance of the GR. This paper seeks to prove their finding from a more general perspective; the asymptotic variance is employed to achieve this goal.

[2] This property also holds for the SR, CN-C, and CN-TR cases.

[3] The diagonal entries are zeros; i.e., $c_{ii} = 0, i = 1, 2, \dots, n$.

2.3 Statistical power and its visualization

Because hypothesis tests are conducted on the basis of samples, they do not always yield correct conclusions. Consequently, considering both Type I (rejecting a true null hypothesis) and Type II (failing to reject a false null hypothesis) errors in hypothesis testing is important. The probability of committing a Type I error is denoted as α, which also is known as the significance level and preset at the beginning of a test procedure; the probability of committing a Type II error is denoted as β, which depends on the sample size, the significance level, and the probability distribution under the null hypothesis. As originally devised, power = $1-\beta$, which is the probability of rejecting a false null hypotheses. The power of a hypothesis test is between zero and one, with a value closer to one indicating a better ability to reject a false null hypothesis. For illustrative purposes, Fig. 11 (Appendix 2) portrays power and is accompanied by description furnishing an intuitive impression about this statistical concept. In a spatial data analyzing procedure, testing for SA in (large) spatial datasets is crucial; accordingly, an obvious question asks about the quality of a test. A formal way or a "standard approach" (Cliff and Ord 1973, p. 131) to evaluate a test is to calculate its statistical power. As Weiss (2017, p. 449) states: "even more helpful is a visual display of the effectiveness of the hypothesis test, obtained by plotting points of power against various values of the parameter and then connecting the points with a smooth curve." This notion of power can be applied to a two-sided (or two-tailed) as well as a one-sided (or one-tailed) situation, in keeping with the type of hypothesis test.[4] For convenience and comparison purposes, Sect. 5 presents the critical values in terms of the MC so that all power curves, including those for the GR and the join counts, can be shown with a single plot. To do so requires establishing theoretical relationship functions linking the MC, the GR, and the join count statistics.

3 Notation and theorems

This section presents necessary notation and limit theorems about variances of the MC and the GR.

Let X be the georeferenced variable of interest distributed over a tessellation. Its observations are $x_1, x_2, \dots, x_n$. The average of these observations is denoted by $\bar{x} = \sum_{i=1}^{n} x_i / n$. $\boldsymbol{C} = \left(c_{ij}\right)_{n \times n}$ is the connectivity matrix denoted in Sect. 2.1.2. The sample MC and GR for variable X are defined as follows:

$$\mathrm{MC} = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{\sum_{i=1}^{n} \sum_{j=1}^{n} c_{ij} \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2}, \tag{1}$$

[4] Given a random variable x, for a two-sided test, power = 1 – probability(x < right critical value) +probability(x > left critical value); for a right-sided test, power = 1 – probability(x < critical value), whereas for a left-sided test, power = probability(x > critical value).

and

$$\mathrm{GR} = \frac{(n-1)\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - x_j\right)^2}{2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}. \tag{2}$$

The GR can be rewritten as (Griffith 1987)

$$\frac{n-1}{2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} \frac{2\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2\left(\sum_{j=1}^{n} c_{ij}\right)}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} - \frac{n-1}{n}\mathrm{MC}. \tag{3}$$

Derivation of this formula appears as proof 1 in Appendix 3.

Cliff and Ord (1973) establish the exact variances of these two statistics. In the following, the subscript *N* denotes normality and *R* denotes randomization:

$$\mathrm{Var}_N(\mathrm{MC}) = \frac{n^2 S_1 - nS_2 + 3S_0^2}{(n-1)(n+1)S_0^2} - \frac{1}{(n-1)^2}, \tag{4}$$

$$\mathrm{Var}_R(\mathrm{MC}) = \frac{n\left[\left(n^2 - 3n + 3\right)S_1 - nS_2 + 3S_0^2\right] - b_2\left[\left(n^2 - n\right)S_1 - 2nS_2 + 6S_0^2\right]}{(n-1)(n-2)(n-3)S_0^2} - \frac{1}{(n-1)^2}, \tag{5}$$

$$\mathrm{Var}_N(\mathrm{GR}) = \frac{\left[\left(2S_1 + S_2\right)(n-1) - 4S_0^2\right]}{2(n+1)S_0^2}, \tag{6}$$

and

$$\begin{aligned}\mathrm{Var}_R(\mathrm{GR}) = {} & \frac{(n-1)S_1\left[n^2 - 3n + 3 - (n-1)b_2\right] - \frac{1}{4}(n-1)S_2\left[n^2 + 3n - 6 - \left(n^2 - n + 2\right)b_2\right]}{n(n-2)(n-3)S_0^2} \\ & + \frac{S_0^2\left[n^2 - 3 - (n-1)^2 b_2\right]}{n(n-2)(n-3)S_0^2},\end{aligned} \tag{7}$$

where $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}$, $S_1 = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(c_{ij} + c_{ji}\right)^2$, $S_2 = \sum_{i=1}^{n}\left[\sum_{j=1}^{n}\left(c_{ij} + c_{ji}\right)\right]^2$, and for $z_i = x_i - \bar{x}$, $b_2 = \frac{1}{n}\sum_{i=1}^{n} z_i^4 / \left(\frac{1}{n}\sum_{i=1}^{n} z_i^2\right)^2$ defines kurtosis. Again, because matrix $\boldsymbol{C}$ is symmetric and binary, $S_1 = 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} = 2S_0$, and $S_2 = 4\sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\right)^2$.

Griffith (2010) proposes simplifying Eqs. (4)–(7) through asymptotics, assuming a normal distribution, producing

$$\mathrm{Var}_A(\mathrm{MC}) = \frac{2}{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}} = \frac{2}{S_0}, \tag{8}$$

and

$$\mathrm{Var}_A(\mathrm{GR}) = \frac{2}{\sum_{i=1}^n \sum_{j=1}^n c_{ij}} + \frac{2\sum_{i=1}^n \left(\sum_{j=1}^n c_{ij}\right)^2}{\left(\sum_{i=1}^n \sum_{j=1}^n c_{ij}\right)^2} = \frac{2}{S_0} + \frac{S_2}{2S_0^2}, \tag{9}$$

where subscript A denotes asymptotic.

Theorems 1 and 2 indicate that the asymptotic variance for the MC is insensitive to swapping the normality and randomization assumptions. They also reveal that the asymptotic variance of the MC approximates the exact variances well for both of these cases.

Theorem 1 $\lim_{n\to\infty} \mathrm{Var}_N(\mathrm{MC}) = \mathrm{Var}_A(\mathrm{MC})$.

Theorem 2 $\lim_{n\to\infty} \mathrm{Var}_R(\mathrm{MC}) = \mathrm{Var}_A(\mathrm{MC})$.

Theorems 3 and 4 discuss the convergence of the GR exact variance for different probability assumptions when sample size approaches infinity. An analogous result has not been obtained for the variance of the GR for permutation sampling; the asymptotic version of this latter index depends on distributional assumptions.

Theorem 3 $\lim_{n\to\infty} \mathrm{Var}_N(\mathrm{GR}) = \mathrm{Var}_A(\mathrm{GR})$.

Theorem 4 $\lim_{n\to\infty} \mathrm{Var}_R(\mathrm{GR})$ depends on b_2, the kurtosis of a distribution.

Proofs for Theorems 1 to 4 appear in Appendix 3. For normal, uniform, beta, and exponential distributions, b_2 has the values 3, 9/5, 3/2, and 9, respectively. Thus,

$$\mathrm{Var}_{\mathrm{AN}}(\mathrm{GR}) = 2/S_0 + S_2/2S_0^2, \tag{10}$$

$$\mathrm{Var}_{\mathrm{AU}}(\mathrm{GR}) = 2/S_0 + S_2/5S_0^2, \tag{11}$$

$$\mathrm{Var}_{\mathrm{AB}}(\mathrm{GR}) = 2/S_0 + S_2/8S_0^2, \quad (\alpha = \beta = 0.5), \tag{12}$$

and

$$\mathrm{Var}_{\mathrm{AE}}(\mathrm{GR}) = 2/S_0 + 2S_2/S_0^2, \tag{13}$$

where the subscripts AN, AU, AB, and AE, respectively, denote the asymptotic variance of the normal, uniform, beta, and exponential distribution. That is to say, the asymptotic variance of the GR is sensitive to distributional assumptions.

Equation (10) coincidences with Griffith's (2010) result [Eq. (9)].

4 Efficiency analysis

This section summarizes results for both asymptotic and exact variances.

4.1 Asymptotic variance ratios

Considering that a statistic with a smaller variance is more efficient, suppose the variance ratio of the MC and the GR is $r_{\text{exact}} = \text{Var}_{\text{exact}}(\text{MC})/\text{Var}_{\text{exact}}(\text{GR})$, where subscript "*exact*" denotes the exact MC and GR variances, given by Eqs. (4) and (6), or by Eqs. (5) and (7). If $r_{\text{exact}} < 1$, then the MC is more efficient than the GR; otherwise, $r_{\text{exact}} > 1$, then the GR is more efficient. The following asymptotic variances also are of interest:

$$r = \text{Var}_A(\text{MC})/\text{Var}_{A*}(\text{GR}) = \frac{2/S_0}{S}, \tag{14}$$

where $A*$ denotes AN, AU, AB, or AE and S denotes Eqs. (10), (11), (12), or (13). Similarly, if $r < 1$, then the MC is more efficient than the GR; if $r > 1$, then the GR is more efficient. Equation (14) indicates that S_0, the sum of ones in matrix $\boldsymbol{C}$, and S_2, the sum of the squared row sums of matrix $\boldsymbol{C}$, are needed to calculate the variance ratio; these two quantities have different values with different geographical configurations, values of S_0 and S_2 of selected geographical configurations are listed in Table 1. More values can be found in Tables 1 and 2 of Luo et al. (2017).

Table 3 presents selected asymptotic variances. Figure 3 portrays their respective ratio curves.

Term k in the last column of Table 3 is the number of constant neighbors; for example, k may be 2, 4, and 8 for the circle, torus rook, and torus queen cases, respectively. Thus, all ratios are less than one, and especially for the maximum planar connectivity case, the values go to zero, which indicates that the MC is more efficient in terms of the asymptotic variances. This property can be seen more clearly from Fig. 3 because its curves have convergent trends whose trajectory values are less than one before n is 100. The CN case is not shown because when k is 2, 4, and 8, its ratios become the same values as those for the L, SR, and SQ cases.

4.2 Exact variance ratios

Section 4.1 presents discussion of asymptotic variance ratios as well as the efficiency priority (i.e., $r < 1$) of the MC versus the GR for the selected probability distributions. One remaining question asks whether or not identical results can be obtained when their exact variances [Eqs. (4) to (7)] are considered. Fortunately, by substituting appropriate S_0 and S_2 values into these formulae, and using some mathematical calculation software (e.g., Wolfram Mathematica 10.2, which has been used for this

Table 3 Asymptotic variance ratios of the MC and the GR

	L	SR	SQ	H	MPC	MH	CN
Normal	1/3	1/5	1/9	1/7	0	3/25	$1/(k+1)$
Uniform	5/9	5/13	5/21	5/17	0	15/59	$5/(2k+5)$
Beta	2/3	½	1/3	2/5	0	6/17	$4/(k+4)$
Exponential	1/9	1/17	1/33	1/25	0	3/91	$1/(4k+1)$

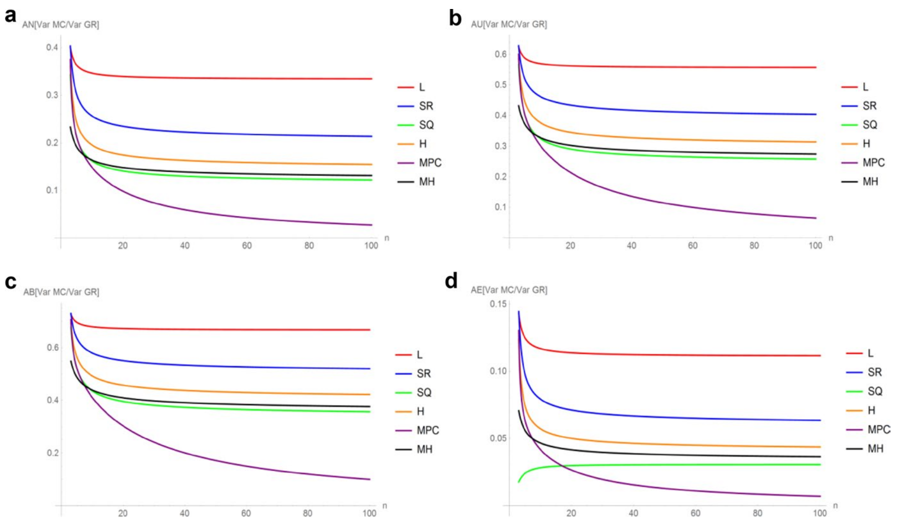

Fig. 3 Asymptotic ratio curves. **a** Curves for normal distribution. **b** Curves for uniform distribution. **c** Curves for beta distribution. **d** Curves for exponential distribution

paper), these exact variance ratios are not difficult to compute. Table 3 in Luo et al. (2017) already summarizes them. That table reveals that all exact variance ratios of the MC versus the GR are one except those for the MPC and MH cases. More specifically, all exact ratios for the MPC case are zero, and they are (approximately) 0.4286, 0.6522, 0.75, and 0.1579 for the MH cases for the selected normal, uniform, beta, and exponential distributions, respectively. These results indicate that the MC is only more efficient than the GR based on the exact variances for the MPC and MH cases. Because the MPC exact ratio is the same as its asymptotic counterpart (both are zero), and because all other asymptotic variance ratios are not one, then these latter asymptotic versions need to be adjusted in order for the ratios to become one. Furthermore, calculating ratios of asymptotic and exact variances of the MC and the GR reveals that the GR's asymptotic variances are the ones that need to be adjusted. The necessary GR adjustment factors equal those exact ratios divided by their asymptotic ratios. For the MPC case, the MC asymptotic variance needs to be adjusted such that it should be multiplied by 1/3 for all probability distributions. This adjustment assessment furnishes quantitative evidence that the GR is far more sensitive to the underlying frequency distribution of an attribute variable.

Luo et al. (2017) present the exact variance ratio curves as well as values for 184 specimen[5] irregular surface partitions. For illustrative and comparative purposes, Fig. 4 reproduces some of these plots more delicately (with a higher resolution and more distinguishable colors), which depict convergence in the

[5] These data come from Griffith (2015); an initial 130 appeared in Griffith (2004), which was expanded to 144 in Griffith and Luhanga (2011).

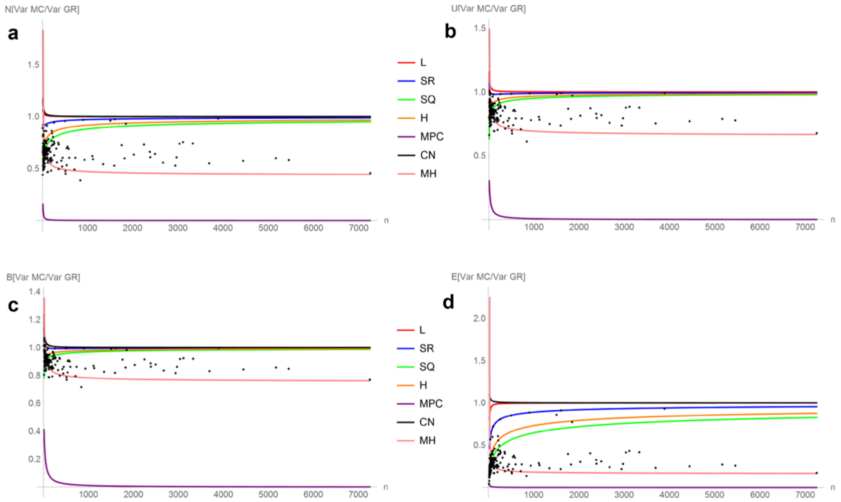

Fig. 4 Exact variance ratio curves with 184 specimen points superimposed. **a** Curves for a normal distribution. **b** Curves for a uniform distribution. **c** Curves for a beta distribution. **d** Curves for an exponential distribution

interval [13, 7250], [10, 7250], [8, 7250], and [23, 7250] for the normal, uniform, beta, and exponential probability distributions, respectively.

Figure 4 portrays that, except for the MPC (the purple curve) ratio converging on zero, and the MH (the pink curve) ratio converging on a specific value less than one for each probability distribution, these ratios converge on one. In addition, those specimen geographic landscapes, presented as black dots superimposed on the ratio curves, mostly scatter between the regular SQ (the green curve) and the MH (the pink curve) cases.

In this section, asymptotic as well as exact variance ratios of the MC and the GR are discussed. These asymptotic variances are far simpler in their expressions than their exact counterparts. This simplicity motivates an exploration of how much the sample size (or threshold) above which those results obtained with asymptotic methods differ from those obtained with exact methods. More detailed work about these two statistics is included in Luo et al. (2017, p. 263, Table 4). One also is interested in these statistics in terms of their asymptotic variance, especially if they have better statistical properties when sample size goes to infinity. For example, for a 1000-by-1000 remotely sensed image, for which the size 1,000,000 far exceeds those thresholds above which asymptotic results are close to exact results (see the square rook row in Table 4 in Luo et al. 2017), both asymptotic variances of the MC and the GR achieve good accuracy.

Consequently, one question asks how to choose between these two indices; this section answers this particular question.

5 Statistical power visualization

Cliff and Ord (1973) conduct simulation experiments to compare the power of the MC and the GR by employing 12-by-2, 4-by-3, 5-by-5, and 7-by-7 lattices in which both the SR and SQ cases are discussed, a 25-cell circle, and the 26 counties of Eire (an irregular surface partitioning), and conclude that the MC is more powerful. Subsequently, they (1981) updated the largest sample size to 81 (a 9-by-9 lattice) by referring to Haining's (1978) work. Being different from the spatial Markov scheme that Cliff and Ord used, Haining introduces a two-dimensional moving average spatial model as the alternative hypothesis, compares the power of the likelihood ratio (denoted by L.R. in his paper) and the MC, and draws the conclusion that the L.R. statistic is more powerful. Writing a year earlier, Bartels and Hordijk (1977) discuss the MC power by using three different error estimators (OLS, BLUS, and RELUS) in their four illustrative examples (the dataset for the first three is the Netherlands with 39 regions, but with a different number of variables for each example, whereas the dataset for the last case is Eire with 26 regions and three artificial variables). All OLS estimators achieve the highest power, except for a very few high (0.9) and low (0.1) SA values. More recently, Dray (2011) develops two new SA indexes to describe a more complex situation (positive and negative SA are involved simultaneously, and their summation is zero or nearly zero), uses a Monte Carlo method to test the significance of these new statistics as well as the MC, and concludes that these two new statistics are as powerful as the MC for purely positive or negative SA structures, but are more powerful than the MC for complex situations.

However, all of these power assessments are calculated based upon Monte Carlo simulations, and only for several selected positive SA values (a one-tailed test). Actually, in the spatial analysis literature, Monte Carlo approaches used for inference are widely adopted not only for areal unit data, but also for point data (Diggle 2010) because of their flexibility, intelligibility, and extendibility. Although Hope (1968) suggests a simplified Monte Carlo test procedure to reduce the size of a reference set (i.e., the number of iterations), generating thousands or even millions of random numbers that are in keeping with a tested distribution still is time-consuming. In addition, this procedure needs to include repetitions. Even reducing the number of iterations by a few in order to reduce the processing time can be at the expense of precision. In contrast, an analytical approach is more rapid and accurate. The following section presents various power curves for the MC and GR with different sample sizes and geographical configurations, which are plotted by an alternative method that appears in Luo et al. (2017).

5.1 A method for calculating statistical power

The definition of statistical power states that if $1 - \beta_{\mathrm{MC}} > 1 - \beta_{\mathrm{GR}}$, then the MC is more powerful than the GR (i.e., the MC test is more likely than the GR test to reject a false null hypothesis, or the MC test is more likely to obtain a significant result to support the existence of a spatially autocorrelated phenomenon); otherwise, the GR is more efficient. Luo et al. (2017) state an alternative hypothesis, H_1, of nonzero SA, which results in two-tailed tests for the MC and GR. But in order to parallel pioneering work and make a clearer comparison, this section focuses on the one-tailed counterpart.[6] Thus, the null hypothesis, H_0, still is no SA, but H_1 becomes a hypothesis of positive SA; here α is set to 0.05.

Figure 5 displays the MC and GR power curves for positive SA and various surface partitionings, where the horizontal axis presents the degree of SA, and the vertical axis stands for the value of statistical power. Each plot shows two sample sizes, 5-by-5 and 9-by-9 (except for the MH with an odd Q, which has two more cells than the other sample sizes), where the solid red–green lines represent the MC-GR power curves with 25 (or 27) cells, and the dashed pink–blue lines represent the MC-GR power curves with 81 (or 83) units. Several conclusions can be made: (1) power increases with increasing sample size (which is a standard result) and the degree of SA; (2) for the SR, SQ, and H cases, the MC is more powerful than the GR; (3) for the L and CN (circle, torus rook, and torus queen) cases, the GR is slightly more powerful than the MC for very small sample sizes (e.g., 5-by-5), but this small advantage disappears with increasing sample size (e.g., 9-by-9); and, (4) for the MH case, the MC is more powerful. Findings (1) to (3) are consistent with Cliff and Ord's (1981) summaries, whereas the MH as well as the H case is newly shown here. Moreover, compared with the early power curves, these Fig. 5 curves are smoother because they are drawn with an analytical method rather than simulation experiments employing only several specific sample sizes; any size power curve can be plotted this way.

5.2 A theoretical evaluation

A key step in the method outlined in Sect. 5.1 is to evaluate the relationship function between the MC and the GR so that their power curves can be plotted with a common measurement scale. In the preceding power analysis, SA is quantified by the MC, so all the GR values are replaced by their respective MC expressions. Fortunately, Eq. (3) furnishes a primary form that indicates a negative correlation between the MC and GR. Referring to this formula, theoretical equations can be constructed. However, in order to construct these equations, the MC and GR values need to be generated. One technique is to take advantage of the matrix $\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{n}\right)\boldsymbol{C}\left(\boldsymbol{I} - \frac{\mathbf{1}\mathbf{1}^{\mathrm{T}}}{n}\right)$,

[6] Following the steps in the mentioned paper, the statistical power of the MC is assessed by replacing all 1.96 values with 1.645 and retaining only the right-hand side of the standardized normal curve. Meanwhile, for the GR, because positive SA is in the interval [0, 1), the one-tailed test is the left-hand side rather than the right-hand side of the standardized normal curve; – 1.96 should be replaced with – 1.645, and the positive portions removed.

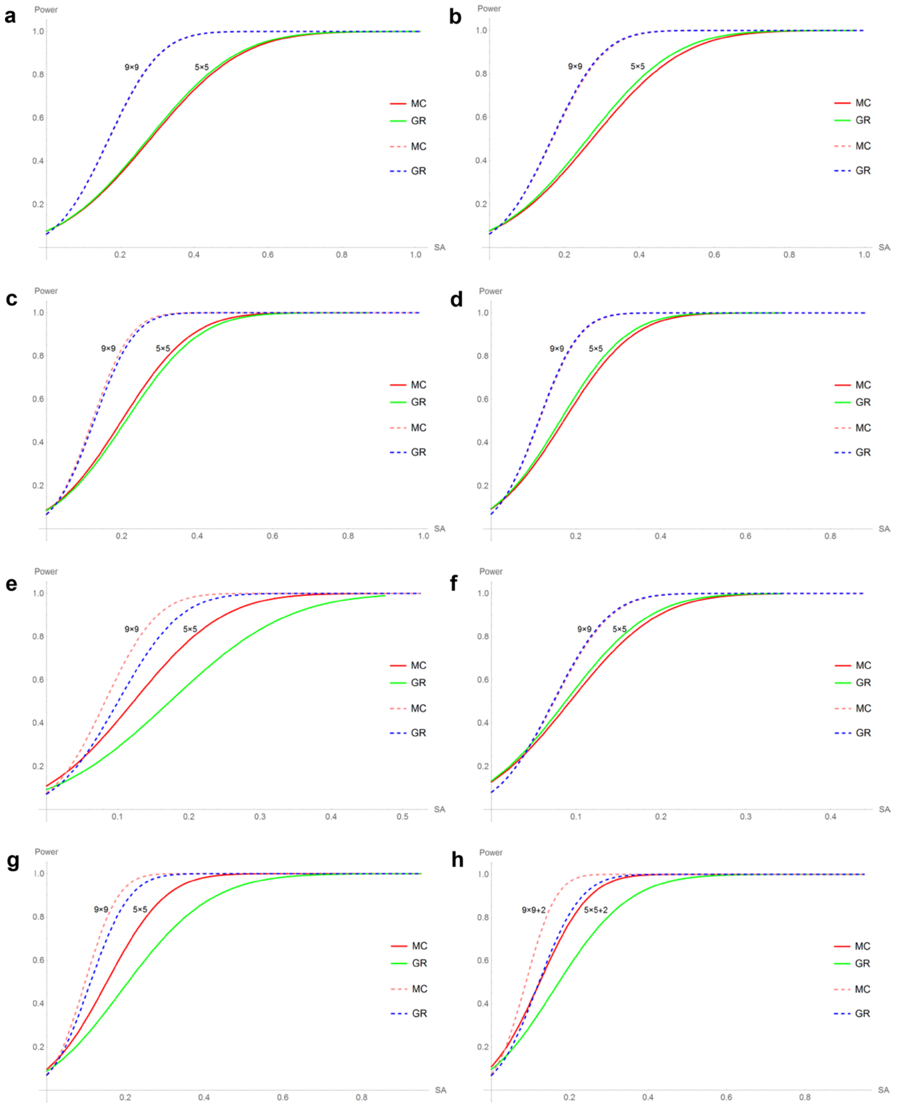

Fig. 5 The MC and GR positive power curves for various geographic configurations. **a** The L case. **b** The CN-C case. **c** The SR case. **d** The CN-TR case. **e** The SQ case. **f** The CN-TQ case. **g** The H case. **h** The MH case

which appears in the numerator of Eq. (1) when the MC is written using matrix notation, where $\boldsymbol{I}$ is the identity matrix, 1 is an n-by-1 vector of ones, and T denotes the matrix transpose operation. Multiplying the eigenvalues of this matrix by $\frac{n}{1^{\mathrm{T}}\boldsymbol{C}1}$ furnishes the complete set of distinct MC values for a geographic landscape, with the extreme values establishing the minimum and maximum possible MC values (de Jong et al. 1984). Corresponding GR values also can be calculated with the eigenvectors of this matrix: using matrix notation, the numerator of Eq. (2) may be written as $2\big((\boldsymbol{C}1)_{\mathbf{diagonal}} - \boldsymbol{C}\big)$ (de Jong et al. 1984; Griffith 2003), where $(\boldsymbol{C}1)_{\mathbf{diagonal}}$ is a diagonal matrix whose diagonal entries are row sums of connectivity matrix $\boldsymbol{C}$. The resulting theoretical relationship functions appear in Luo et al. (2017).

Figure 6 portrays selected scatter plots with fitted lines (shown in red) superimposed on them. These plots depict the relationship between the GR (the vertical axis) and the MC (the horizontal axis) for regular tessellations (the SR, SQ, and H cases for $n = 10,000$) and the CN case. Overall, these scatter plots have a negative sloping trend line, although thick line portions appear in the SR, SQ, and H scatter plots. All of the fitted lines evaluated by the functions that have the same form as Eq. (3) closely correspond to their respective scatter plots.

6 The MC and GR versus the join count statistics

As one type of test for SA, the join count statistics (Cliff and Ord 1973) apply to nominal (e.g., binary 0–1) data. Three different join count statistics exist: BB, WW, and BW, where BB denotes a one area adjacent to a one area, WW denotes a zero area adjacent to a zero area, and BW denotes a one adjacent to a zero area. Assignment of the value one or zero to the ith areal unit depends on the presence

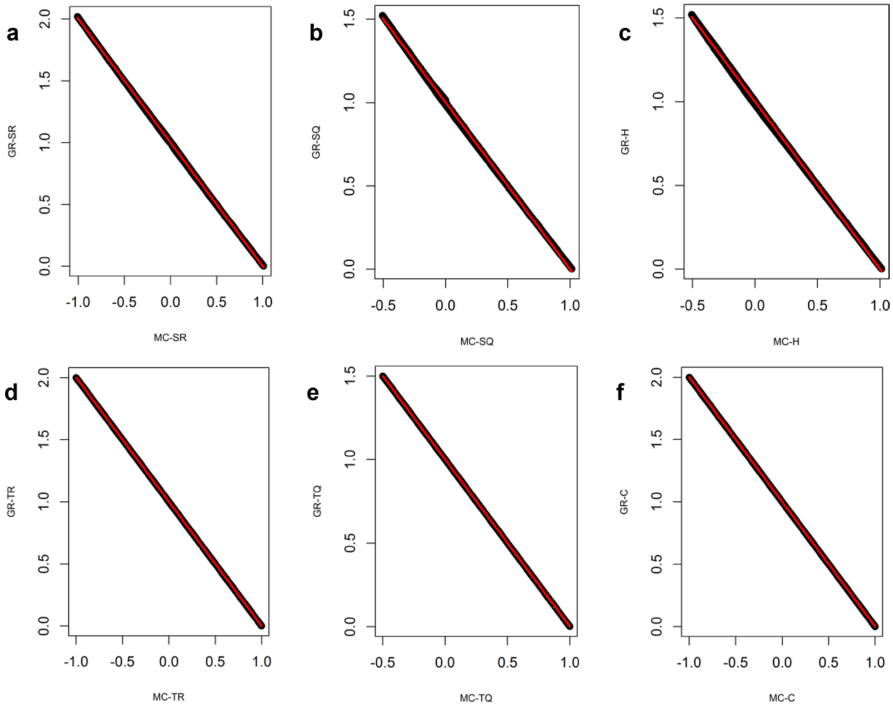

Fig. 6 MC versus GR scatter plots with superimposed fitted lines. **a** The SR case. **b** The SQ case. **c** The H case. **d** The CN-TR case. **e** The CN-TQ case. **f** The CN-C case

or absence of some phenomenon in that unit. If it is present, then this unit has $x_i = 1$; otherwise, it has $x_i = 0$.

Cliff and Ord (1973) point out the similarity of the BB and MC, and the BW and GR, furnish an equation relating the BB and MC in which the attribute variable X also is included, and derive WW as a linear combination of BB and BW. Although these join count statistics are less popular today than several decades ago, Chun and Griffith (2013) furnish equations relating the MC and BB + WW, and the GR and BW for nonfree sampling (sampling without replacement):

$$\mathrm{MC} = \frac{2n}{S_0}\left(\frac{\mathrm{BB}}{n_1} + \frac{\mathrm{WW}}{n_2}\right) - 1, \tag{15}$$

and

$$\mathrm{GR} = \frac{n(n-1)}{S_0}\frac{\mathrm{BW}}{n_1 n_2}, \tag{16}$$

where n_1 is the number of areal units with one, n_2 is the number areal units with zero, n_1 and n_2 are preset, and $n_1 + n_2 = n$.

On one hand, Eq. (16) confirms the similarity between the GR and BW; on the other hand, Eq. (15) indicates that the MC is related not only to the BB but also to the WW. Cliff and Ord (1973) only considered the similarity between the MC and BB. Because WW can be written as a linear combination of BB and BW, the MC finally relates to BB and BW.

Again, using the technique suggested in Sect. 5, Fig. 7 portrays selected GR and BW power plots of two-tailed tests for the CN-C, SR, SQ, and H cases and the 5-by-5 and 9-by-9 sample sizes. The solid red–green lines represent 5-by-5 GR-BW

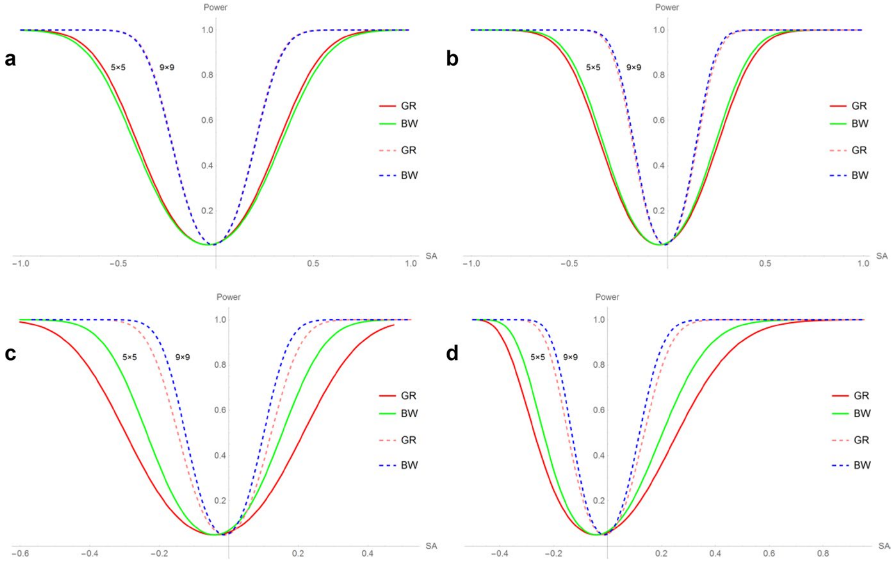

Fig. 7 The two-tailed test power plots of the GR versus the BW. **a** The CN-C case. **b** The SR case. **c** The SQ case. **d** The H case

power plots, whereas the dashed pink–blue lines represent the 9-by-9 GR-BW power curves. Except the CN-C case, all plots depict a power priority of the BW versus the GR, which is counter to Cliff and Ord's (1973) results.

7 Two massive spatial data examples

Two remotely sensed images are employed to verify the findings furnished in the previous sections. For the continuous random variables, the normalized difference vegetation index (NDVI) was calculated for a Landsat 7 Enhanced Thematic Mapper Plus (ETM+) image of the Yellow Mountain region (Anhui, China) to illustrate the efficiency of the MC versus the GR. To illustrate a nominal data case, pixels constituting an image of the Huairou Reservoir region (Beijing, China) captured from Map World are classified as water or not water to calculate the join count test as well as to indicate weaknesses of the GR versus the MC for this measurement scale.

7.1 A continuous random variable case

A Yellow Mountain image (Fig. 8), downloaded from the USGS Earth Explore website (https://earthexplorer.usgs.gov/), is for October 8, 2002, and forms a 7811-by-7051 rectangular region with $n = 55,075,361$ pixels. It includes spectral bands B1–B8, with B1–B7 having 30 m spatial resolution, and B8 having 15 m spatial resolution. Considering there are some zero spectral value regions black areas in Fig. 8a), and the borders are indented, a 5140-by-4754 ($n = 24,435,560$) pixels sub-image (Fig. 8b) was cropped and is the study area across which the NDVI is calculated. Figure 8b is the zoomed-in version of the area demarcated by the red border in Fig. 8a.

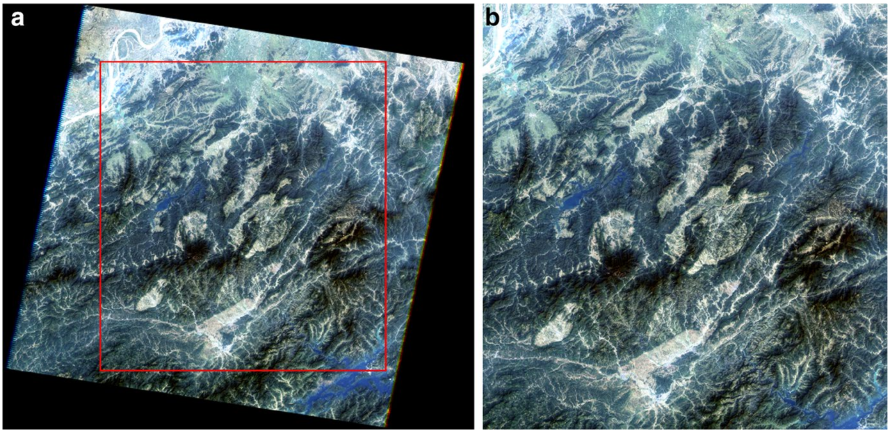

Fig. 8 The Yellow Mountain region remotely sensed image and the subarea extracted for analysis

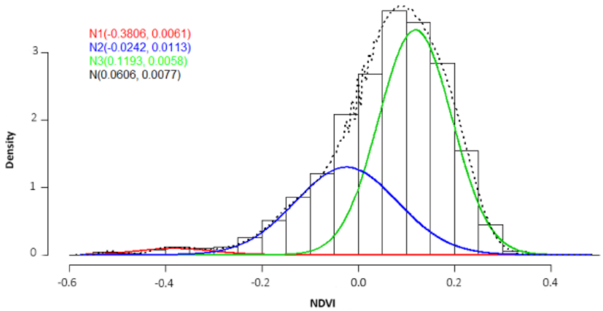

Fig. 9 Normal finite mixture distribution with three components for the Yellow Mountain subregion

Table 4 Selected statistics for the NDVI of the Yellow Mountain region sub-image

	Value	Expected value	Variance	Z score	Asymptotic variance	Exact variance ratio	Asymptotic variance ratio	Power
MC	0.9294	−4.0924e−8	2.0466e−8	6496.4809	2.0466e−8	0.9999	0.2000	1
GR	0.0705	1	2.0470e−8	6496.7667	1.0232e−7			1

The distribution of the NDVI is shown in Fig. 9; three normal distributions (denoted by red, green, and blue curves) are fitted to these data as components of a finite mixture distribution (black dotted line curve).

Table 4 includes the MC and GR values as well as some hypothesis testing statistics obtained with the normality assumption for the NDVI index. By setting rook adjacency and constructing the binary spatial weights matrix, the values of the MC and the GR are 0.9294 and 0.0705, respectively, which indicate very strong positive SA. Meanwhile, the expected values, variances, and Z-scores under the null hypothesis of zero SA are listed; the extremely large Z-scores imply rejection of the null hypothesis. The asymptotic variances as well as their ratio of 0.2 (this calculated value coincides with the theoretically derived value; see the entry of the normal row and the SR column in Table 3) support an efficiency priority for the MC versus the GR. The power values go to one because of the large sample size, which is 24,435,560 here.

7.2 A binary random variable case

The Huairou Reservoir region image is captured from Map World (http://www.tianditu.cn/). It was obtained in the summer of 2010 by ZY-3 and covers a 6843-by-7895 rectangle area comprising $n = 54,025,485$ pixels; it is an RGB image. However, for analysis purposes, this image is dichotomized, with pixels being classified as being water or not water. Specifically, those values of pixels with no water are set to 0 ($n_2 = 45,977,103$; 85.10% of the total), and those values of pixels with water are set one ($n_1 = 8,048,382$; 14.90% of the total). Figure 10 portrays the original image and its binary counterpart.

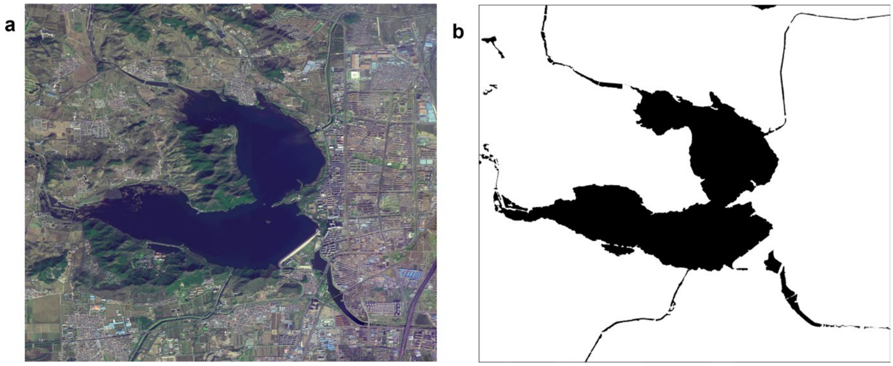

Fig. 10 The Huairou Reservoir remotely sensed image and its binary counterpart

Table 5 Selected statistics for the dichotomized Huairou Reservoir image

	Value	Expected value	Variance	Z score	Asymptotic variance	Exact variance ratio	Asymptotic variance ratio	Power
MC	0.9969	−1.8510e−8	9.2562e−9	1.0362e+4	9.2562e−9	0.9999	0.2000	1
GR	0.0032	1	9.2574e−9	−1.0361e+4	4.6277e−8			1
BB	16,052,502	2,397,669	1,736,584	1.0362e+4	–	–	–	1
WW	91,895,211	78,244,764	1,739,205	1.0351e+4	–	–	–	–
BW	88,519	27,393,799	6,947,846	−1.0359e+4	–	–	–	1

Table 5 summarizes results for statistical hypothesis tests conducted in terms of the join count statistics, the MC, and the GR for this binary image. Hypothesis testing with the join count statistics is under nonfree sampling, whereas hypothesis testing with the other two statistics is under normality; all three utilize the rook adjacency. Statistics in Table 5 imply the present of significant positive SA because the counts of BB and WW joins are larger than their respective expectations, and the BW join count is significantly less than its expectation, both of which confirm the rejection of zero SA. Meanwhile, the MC and GR values are very close to their extreme positive values. In addition, the asymptotic variance ratio for the MC versus the GR also indicates a weakness of the GR for this large sample size. As an aside, these quantities together with the n_1 and n_2 values confirm Eqs. (15) and (16). The large sample size of 54,025,485 produces statistical powers of one.

8 Conclusions and discussions

In its formative years, spatial statistics restricted much of its attention to small-to-medium datasets mostly because of computer technology constraints; more recently, it commonly engages large-to-massive datasets because computer technology allows it to. Therefore, analyses of properties of SA statistics for massive spatial data are necessary. This paper focuses on the efficiency and statistical power of the MC and the GR for massively large sample sizes and draws two main conclusions. Firstly, the MC is more efficient than the GR in terms of asymptotic variances, but only for the MPC and the MH cases when exact variances are discussed. (The MC and the GR can be equally efficient for other geographical configurations.) This is a finding that alters our understanding of the MC and the GR.

Secondly, the statistical power of these two indexes goes to one when sample size is large, negating some results established with small datasets. A number of additional findings also are important. One is that the asymptotic variance of the MC is more stable across, and hence less sensitive to, distributional assumptions, a conclusion implied by Theorems 1 to 4, because the asymptotic variance of the MC may be uniformly expressed by the formula introduced by Griffith (2010), whereas the one for the GR is determined by an underlying distribution's kurtosis. The second finding is that the relative efficiency positions of the 184 empirical irregular surface partitioning specimens indicate that realistic geographic surface partitionings are between the MH and the regular SR or SQ configurations. The third finding is that the relationship between the MC and the GR may be expressed by Eq. (3), which highlights a negative correlation between these two statistics, and allows them to be differentiated according to attribute variable and connectivity features. A final finding complements results obtained by Cliff and Ord (1973): the MC is not more powerful than the GR for all possible geographical configuration types (e.g., the L and CN cases) and relatively large sample sizes. These asymptotic variance, efficiency, and power comparison results for large sample sizes and various spatial structures are relevant to especially massive spatial data analyses.

In addition, a comparative power visualization technique is presented in this paper that produces smoother power curves for any sample size. Plots appearing in Fig. 5 are generated by this technique; they contain more connectivity cases than those presented in Luo et al. (2017) and reveal that the MC is not more powerful than the GR for positive SA when the connectivity criteria are L or CN. Instead of obtaining the *p* value by ranking the test criteria with those random sample results (Hope 1968), this technique calculates the probability and the power through a formal inference protocol, and its significant results can lead to a rejection of the null hypothesis of zero SA, whereas the significant results of a Monte Carlo test can only indicate no spatial randomness. Finally, a discussion of the join count statistics reveals that Cliff and Ord (1973) might have focused on BB + BW rather than only on BB when considering the similarity between the MC and the join count statistics. The GR-BW power plots appearing in Fig. 7 reveal a surprising conclusion that the BW is more powerful than the GR for the SR, SQ, and H cases.

Once again, the conclusion is that the MC is preferable to the GR for a big spatial data analysis that always contains massive samples and has complex geographical configurations because the asymptotic variance of the former is smaller and more stable than that of the latter. Furthermore, the statistical power of these two statistics as well as the join count statistics approach one under the situation of big spatial data, i.e., the power advantage of any statistic existing in small samples, is lost.

In conclusion, this paper focuses on statistical properties of two SA coefficients in the background of big spatial data. These statistics need to be calculated no matter how big a sample size is–they emphasize different features of a spatial dataset and furnish input for choosing a proper model specification, which is a different issue from the meaningless statistical significance that arises from a massive sample size. The discussion of SA throughout this paper is from a global perspective; a local perspective that may relate to spatial heteroskedasticity also is relevant sometimes. This spatial heteroskedasticity refers to unstable/different means, variances, and possibly frequency distributions across a geographical landscape. Spatially varying means can be described with regression covariates. Spatially varying variances can be adjusted for along the lines of Oden (1995), Waldhör (1996), and Jackson et al. (2010); this is a future research topic. Griffith and Chun (2016) address yet another aspect of varying variance, namely the uncertainty of the SA parameter in a simultaneous autoregressive (SAR) model; they describe it with a beta–beta mixture distribution. Spatially varying frequency distributions can be assessed with diagnostic statistics (one goal here is sensitivity to specification error). Spatially varying SA can generate an outcome of different levels of local SA [i.e., LISA; (e.g., local Moran's I), whose linear combination is proportional to the global SA coefficient]. Consequently, one manifestation for a given geographic landscape is that many significant LISA average to approximately zero, implying their global SA measure is not significant. Spatial heteroskedasticity reflects the uneven changes of geographical phenomena, changes that may be attributable to geographic diversity. Thus, spatial heteroskedasticity is a symptom of an inhomogeneous geographical landscape. Transcending this local perspective, SA also can be discussed in terms of a model perspective, because the MC can express the SA parameter rho in a SAR model, for example, as a sigmoid function (Griffith, 2003, p. 33); this link function provides a connection between the MC and spatial autoregressive models that contain many meaningful covariates. In this context, the SA term including rho in an SAR model represents missing spatially structured variables, hence substituting for uninclude covariates. These themes constitute topics for future research.

Acknowledgements Funding was provided by The National Key Research and Development Program of China (Grant No. 2017YFB0503802) and China Scholarship Council (Grant No. 201406270075).

Appendix 1: Selected eigenvalues of binary connectivity matrices and corresponding MC and GR values for three theoretical configurations

See Table 6.

Table 6 First and last ten values of λ, MC, and GR

Configuration type	n	values	Rank order of eigenvalues (integers alone denote descending; integers in parentheses denote ascending)									
			1	2	3	4	5	6	7	8	9	10
			(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
MPC	102	λ	15.641	1.9961	1.9921	1.9845	1.9766	1.9653	1.9536	1.9384	1.9229	1.9040
			−1.9221	−1.9384	−1.9528	−1.9653	−1.9759	−1.9845	−1.9913	−1.9961	−1.9990	−12.661
		MC	0.3393	0.3387	0.3374	0.3360	0.3341	0.3321	0.3295	0.3269	0.3237	0.3204
			−0.3268	−0.3295	−0.3320	−0.3341	−0.3359	−0.3374	−0.3385	−0.3393	−0.3398	−0.4934
		GR	0.3373	0.3379	0.3392	0.3404	0.3423	0.3442	0.3466	0.3491	0.3522	0.3553
			0.9992	1.0017	1.0039	1.0058	1.0073	1.0085	1.0093	1.0098	17.169	17.17
	10,002	λ	142.92	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000	2.0000
			−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−2.0000	−139.92
		MC	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334
			−0.3334	−0.3334	−0.3334	−0.3334	−0.3334	−0.3334	−0.3334	−0.3334	−0.3334	−0.4999
		GR	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334	0.3334
			1.0001	1.0001	1.0001	1.0001	1.0001	1.0001	1.0001	1.0001	1667.2	1667.2

Table 6 (continued)

Configuration type	n	values	Rank order of eigenvalues (integers alone denote descending; integers in parentheses denote ascending)									
			1	2	3	4	5	6	7	8	9	10
			(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
MH-O	112	λ	6.8854	6.0328	5.5520	5.3644	4.9839	4.8983	4.7366	4.3765	4.1273	4.1111
			−2.4743	−2.4886	−2.6745	−2.6763	−2.6877	−2.7006	−2.8373	−2.8757	−3.7881	−4.3096
		MC	1.0237	1.0095	0.9103	0.8457	0.8312	0.8041	0.7427	0.7106	0.7004	0.6781
			−0.4199	−0.4223	−0.4535	−0.4542	−0.4561	−0.4566	−0.4807	−0.4879	−0.6026	−0.7313
		GR	0.0971	0.1481	0.1897	0.2368	0.2815	0.3105	0.3197	0.3303	0.3413	0.5045
			1.3997	1.4025	1.4425	1.4441	1.4528	1.4593	1.4848	1.5053	2.5284	3.1188
	10,102	λ	15.890	14.845	5.9965	5.9923	5.9907	5.9866	5.9847	5.9811	5.9789	5.9771
			−2.9933	−2.9933	−2.9955	−2.9955	−2.9961	−2.9962	−2.9983	−2.9983	−12.878	−13.785
		MC	2.5971	2.4746	0.9989	0.9986	0.9980	0.9979	0.9973	0.9967	0.9964	0.9960
			−0.4990	−0.4990	−0.4993	−0.4993	−0.4995	−0.4995	−0.4998	−0.4998	−2.1080	−2.2980
		GR	0.0013	0.0016	0.0022	0.0024	0.0030	0.0036	0.0038	0.0042	0.0050	0.0051
			1.4994	1.4994	1.4995	1.4995	1.4999	1.4999	12.549	13.549	19.518	20.864

Table 6 (continued)

Configuration type	*n*	values	Rank order of eigenvalues (integers alone denote descending; integers in parentheses denote ascending)									
			1	2	3	4	5	6	7	8	9	10
			(10)	(9)	(8)	(7)	(6)	(5)	(4)	(3)	(2)	(1)
MH-E	102	λ	6.8165	5.9401	5.4932	5.3392	4.8887	4.7451	4.6165	4.1772	4.0048	3.9292
			−2.4369	−2.4461	−2.6359	−2.6428	−2.6602	−2.6602	−2.8319	−2.8636	−3.7526	−4.1866
		MC	1.0098	0.9921	0.9077	0.8311	0.8067	0.7848	0.7101	0.6808	0.6808	0.6648
			−0.4143	−0.4158	−0.4481	−0.4493	−0.4495	−0.4523	−0.4812	−0.4868	−0.6043	−0.7117
		GR	0.1016	0.1596	0.2165	0.2641	0.2930	0.3037	0.3315	0.3544	0.3624	0.3646
			1.3880	1.4279	1.4339	1.4344	1.4567	1.4598	1.4841	1.4915	2.4404	3.0089
	10,002	λ	15.855	14.810	5.9965	5.9922	5.9905	5.9864	5.9847	5.9808	5.9788	5.9768
			−2.9932	−2.9932	−2.9954	−2.9954	−2.9961	−2.9961	−2.9983	−2.9983	−12.842	−13.750
		MC	2.5910	2.4688	0.9989	0.9986	0.9979	0.9979	0.9972	0.9967	0.9963	0.9960
			−0.4990	−0.4990	−0.4993	−0.4993	−0.4995	−0.4995	−0.4998	−0.4998	−2.1018	−2.2922
		GR	0.0013	0.0017	0.0022	0.0024	0.0030	0.0036	0.0038	0.0042	0.0050	0,0051
			1.4994	1.4994	1.4995	1.4995	1.4999	1.4999	12.472	13.474	19.430	20.771

Some of the values are the same in this table because they are round due to space limitations. The red figures are extremely large eigenvalues of matrix ***C***, and the corresponding MC and GR values

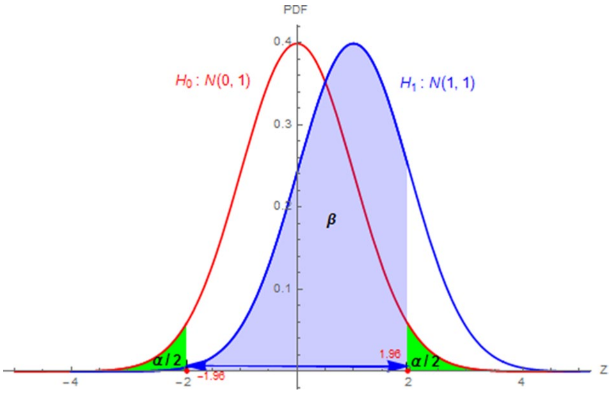

Fig. 11 An example of hypothesis testing

Appendix 2: A descriptive introduction of statistical power

Figure 11 shows necessary elements of a hypothesis testing procedure. Suppose one is testing the null hypothesis *mean*=*0* whose underlying distribution is standard normal, setting the significance level α to 0.05, which results in the critical values ± 1.96. Suppose the true mean value is one, which is the alternative hypothesis. The two green areas are critical regions in which the null hypothesis will be rejected; thus, the interval [− 1.96, 1.96] is the range across which the null will not be rejected. Because the true mean is one, failing to reject the null commits a Type II error, which is the area colored blue under the alternative distribution curve (the blue normal curve). Therefore, the statistical power of this hypothesis testing example is the areas under the blue curve that are restricted to $[1.96, +\infty)$ and $(-\infty, -1.96]$.

Appendix 3: Proofs for the relationship function between the MC and the GR and Theorems 1 to 4

Proof 1 Substituting Eq. (1) into Eq. (3) yields

$$\frac{(n-1)\left[2\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2\left(\sum_{j=1}^{n}c_{ij}\right)-2\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\left(x_i-\bar{x}\right)\left(x_j-\bar{x}\right)\right]}{2\sum_{i=1}^{n}\sum_{j=1}^{n}c_{ij}\sum_{i=1}^{n}\left(x_i-\bar{x}\right)^2}.$$

Comparing this equation to Eq. (2), the proof requires only showing the equality of their numerators. Considering $\left(x_i-x_j\right)^2=\left[\left(x_i-\bar{x}\right)-\left(x_j-\bar{x}\right)\right]^2$, and utilizing the symmetry of matrix $\boldsymbol{C}$, yields

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - x_j\right)^2 \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_j - \bar{x}\right)^2 \\
&= 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) \\
&= 2\sum_{i=1}^{n} \left(c_{i1} + c_{i2} + \cdots + c_{in}\right)\left(x_i - \bar{x}\right)^2 - 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right) \\
&= 2\sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2\left(\sum_{j=1}^{n} c_{ij}\right) - 2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right).
\end{aligned}$$

$\therefore GR =$ Eq. (3). □

The following are proofs for Theorems 1 to 4 (T1 to T4).

Proof of T1

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}_N(\mathrm{MC}) \\
&= \lim_{n\to\infty} \frac{n^2(n-1)S_1 - n(n-1)S_2 + 3(n-1)S_0^2 - (n+1)S_0^2}{(n-1)^2(n+1)S_0^2} \\
&= \lim_{n\to\infty} \left[\frac{n^2 S_1}{\left(n^2-1\right)S_0^2} - \frac{nS_2}{\left(n^2-1\right)S_0^2} + \frac{2(n-2)}{(n-1)^2(n+1)}\right] \\
&= \frac{S_1}{S_0^2} - o(1)\frac{S_2}{S_0^2} + 2o\left(\frac{1}{n}\right) = \frac{2}{S_0} = \mathrm{Var}_A(\mathrm{MC}),
\end{aligned}$$

where $o(1) = 1/n$ is an infinitesimal over $n \to \infty$, S_2/S_0^2 is a constant (it is a positive constant for the maximum planar connectivity case; otherwise, it converges to zero), and $o(1/n) = 1/n^2$ is the infinitesimal of higher order than $1/n$ over $n \to \infty$. □

Proof of T2

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}_R(\mathrm{MC})\\
&= \lim_{n\to\infty}\left\{ \frac{n(n-1)\left[\left(n^2-3n+3\right)S_1 - nS_2 + 3S_0^2\right] - b_2(n-1)\left[\left(n^2-n\right)S_1 - 2nS_2 + 6S_0^2\right]}{(n-1)^2(n-2)(n-3)S_0^2}\right.\\
&\quad \left. - \frac{(n-2)(n-3)S_0^2}{(n-1)^2(n-2)(n-3)S_0^2}\right\}\\
&= \lim_{n\to\infty}\left\{ \frac{n\left(n^2-3n+3\right)S_1}{(n-1)(n-2)(n-3)S_0^2} - \frac{n^2S_2}{(n-1)(n-2)(n-3)S_0^2} + \frac{3n}{(n-1)(n-2)(n-3)}\right.\\
&\quad \left. -b_2\left[\frac{nS_1}{(n-2)(n-3)S_0^2} - \frac{2nS_2}{(n-1)(n-2)(n-3)S_0^2} + \frac{6}{(n-1)(n-2)(n-3)}\right] - \frac{1}{(n-1)^2}\right\}\\
&= \frac{S_1}{S_0^2} - o(1)\frac{S_2}{S_0^2} + 3o\left(\frac{1}{n}\right)\\
&\quad - b_2\left[o(1)\frac{S_1}{S_0^2} - 2o\left(\frac{1}{n}\right)\frac{S_2}{S_0^2} + 6o\left(\frac{1}{n^2}\right)\right] - o\left(\frac{1}{n}\right)\\
&= \frac{S_1}{S_0^2} = \frac{2}{S_0} = \mathrm{Var}_A(\mathrm{MC}),
\end{aligned}
$$

where b_2 is a constant (an index of kurtosis) whose value may vary with the assumed distribution, and $o\left(1/n^i\right)(i = 0, 1, 2)$ are infinitesimals (of higher order) over $n \to \infty$. □

Proof of T3

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}_N(\mathrm{GR})\\
&= \lim_{n\to\infty}\left[\frac{\left(2S_1+S_2\right)(n-1)}{2(n+1)S_0^2} - \frac{2}{(n+1)}\right]\\
&= \frac{\left(2S_1+S_2\right)}{2S_0^2} - 2o(1) = \frac{2}{S_0} + \frac{S_2}{2S_0^2}.
\end{aligned}
$$

$\therefore \lim_{n\to\infty} \mathrm{Var}_N(\mathrm{GR}) = \mathrm{Var}_A(\mathrm{GR})$ □

Proof of T4

$$
\begin{aligned}
&\lim_{n\to\infty} \mathrm{Var}_R(\mathrm{GR}) \\
&= \lim_{n\to\infty} \left\{ \frac{(n-1)S_1\left[n^2-3n+3-(n-1)b_2\right] - \frac{1}{4}(n-1)S_2\left[n^2+3n-6-\left(n^2-n+2\right)b_2\right]}{n(n-2)(n-3)S_0^2} \right. \\
&\quad \left. + \frac{S_0^2\left[n^2-3-(n-1)^2 b_2\right]}{n(n-2)(n-3)S_0^2} \right\} \\
&= \lim_{n\to\infty} \left[\frac{(n-1)\left(n^2-3n+3\right)S_1}{n(n-2)(n-3)S_0^2} - \frac{(n-1)^2 S_1 b_2}{n(n-2)(n-3)S_0^2} - \frac{(n-1)\left(n^2+3n-6\right)S_2}{4n(n-2)(n-3)S_0^2} \right. \\
&\quad \left. + \frac{(n-1)\left(n^2-n+2\right)S_2 b_2}{4n(n-2)(n-3)S_0^2} + \frac{n^2-3}{n(n-2)(n-3)} - \frac{(n-1)^2 b_2}{n(n-2)(n-3)} \right] \\
&= \frac{S_1}{S_0^2} - o(1)\frac{S_1}{S_0^2}b_2 - \frac{S_2}{4S_0^2} + \frac{S_2 b_2}{4S_0^2} + o(1) - o\left(\frac{1}{n}\right)b_2 \\
&= \frac{2}{S_0} + \frac{S_2\left(b_2-1\right)}{4S_0^2}.
\end{aligned}
$$

□

References

Anselin L (1995) Local indicators of spatial association—LISA. Geogr Anal 27(2):93–115. https://doi.org/10.1111/j.1538-4632.1995.tb00338.x

Anselin L (1996) The Moran scatterplot as an ESDA tool to assess local instability in spatial association. In: Fischer M, Scholten H, Unwin D (eds) Spatial analytical perspectives on GIS. Taylor and Francis, London, pp 111–125

Anselin L (2018) A local indicator of multivariate spatial association: Extending Geary's c. Geogr Anal. https://doi.org/10.1111/gean.12164

Bartels CPA, Hordijk L (1977) On the power of the generalized Moran contiguity coefficient in testing for spatial autocorrelation among regression distributions. Reg Sci Urban Econ 7(1):83–101. https://doi.org/10.1016/0166-0462(77)90019-9

Bavaud F (2013) Testing spatial autocorrelation in weighted networks: The modes permutation test. J Geogr Syst 15(3):233–247. https://doi.org/10.1007/s10109-013-0179-2

Bivand R, Müller WG, Reder M (2009) Power calculations for global and local Moran's I. Comput Stat Data Anal 53(8):2859–2872. https://doi.org/10.1016/j.csda.2008.07.021

Boots B (2003) Developing local measure of spatial association for categorical data. J Geogr Syst 5(2):139–160. https://doi.org/10.1007/s10109-003-0110-3

Boots B, Tiefelsdorf M (2000) Global and local spatial autocorrelation in bounded regular tessellations. J Geogr Syst 2(4):319–348. https://doi.org/10.1007/PL00011461

Carrijo TB, da Silva AR (2017) Modified Moran's I for small samples. Geogr Anal 49(4):451–467. https://doi.org/10.1111/gean.12130

Cheng T, Haworth J, Wang J (2012) Spatio-temporal autocorrelation of road network data. J Geogr Syst 14(4):389–413. https://doi.org/10.1007/s10109-011-0149-5

Chun Y (2008) Modeling network autocorrelation within migration flows by eigenvector spatial filtering. J Geogr Syst 10(4):317–344. https://doi.org/10.1007/s10109-008-0068-2

Chun Y, Griffith DA (2013) Spatial statistics and geostatistics: theory and applications for geographic information science and technology. SAGE, Thousand Oaks

Cliff AD, Ord JK (1969) The problem of spatial autocorrelation. In: Scott AJ (ed) Studies in regional science. Pion Ltd, London, pp 25–55

Cliff AD, Ord JK (1970) Spatial autocorrelation: A review of existing and new measures with applications. Econ Geogr 46:269–292. https://doi.org/10.2307/143144

Cliff AD, Ord JK (1973) Spatial autocorrelation. Pion Ltd, London

Cliff AD, Ord JK (1981) Spatial process. Pion Ltd, London

de Jong P, Sprenger C, van Veen F (1984) On extreme values of Moran's I and Geary's c. Geogr Anal 16(1):17–24. https://doi.org/10.1111/j.1538-4632.1984.tb00797.x

de la Mata T, Llano C (2013) Social networks and trade of service: Modelling interregional flows with spatial and network autocorrelation. J Geogr Syst 15(3):319–367. https://doi.org/10.1007/s10109-013-0183-6

Diggle P (2010) Nonparametric methods. In: Gelfand AE, Diggle PJ, Fuentes M, Guttorp P (eds) Handbook of spatial statistics. CRC Press, Baca Raton, pp 299–316

Dray S (2011) A new perspective about Moran's Coefficient: Spatial autocorrelation as a linear regression problem. Geogr Anal 43(2):127–141. https://doi.org/10.1111/j.1538-4632.2011.00811.x

Geary RC (1954) The contiguity ratio and statistical mapping. Inc Stat 5(3):115–146. https://doi.org/10.2307/2986645

Griffith DA (1987) Spatial autocorrelation: a primer. AAG, Pennsylvania

Griffith DA (1996) Spatial autocorrelation and eigenfunctions of the geographic weights matrix accompanying geo-referenced data. Can Geogr 40(4):351–367. https://doi.org/10.1111/j.1541-0064.1996.tb00462.x

Griffith DA (2003) Spatial autocorrelation and spatial filtering: gaining understanding through theory and scientific visualization. Springer, Berlin

Griffith DA (2004) Extreme eigenfunctions of adjacency matrices for planar graphs employed in spatial analyses. Linear Algebra Appl 388:201–219. https://doi.org/10.1016/S0024-3795(03)00368-9

Griffith DA (2010) The Moran coefficient for non-normal data. J Stat Plan Inference 140(11):2980–2990. https://doi.org/10.1016/j.jspi.2010.03.045

Griffith DA (2015) On the eigenvalue distribution of adjacency matrices for connected planar graphs. Quaest Geogr. https://doi.org/10.1515/quageo-2015-0035

Griffith D, Chun Y (2016) Spatial autocorrelation and uncertainty associated with remotely-sensed data. Remote Sens 8(7):535. https://doi.org/10.3390/rs8070535

Griffith DA, Luhanga U (2011) Approximating the inertia of the adjacency matrix of a connected planar graph that is the dual of a geographic surface partitioning. Geogr Anal 43(4):383–402. https://doi.org/10.1111/j.1538-4632.2011.00828.x

Haining RP (1978) The moving average model for spatial interaction. Trans Inst Br Geogr 3(2):202–225. https://doi.org/10.2307/622202

Haynes D, Jokela A, Manson S (2018) IPUMS-Terra: Integrated big heterogeneous spatiotemporal data analysis system. J Geogr Syst 20(4):343–361. https://doi.org/10.1007/s10109-018-0277-2

Hope ACA (1968) A simplified Monte Carlo significance test procedure. J R Stat Soc B 30(3):582–598

Jackson MC, Huang L, Xie Q, Tiwari RC (2010) A modified version of Moran's I. Int J Health Geogr 9:33. https://doi.org/10.1186/1476-072X-9-33

Lee SI (2001) Developing a bivariate spatial association measure: an integration of Pearson's r and Moran's I. J Geogr Syst 3(4):369–385. https://doi.org/10.1007/s101090100064

Lee J, Kang M (2015) Geospatial big data: challenges and oppurtunities. Big Data Res 2(2):74–81. https://doi.org/10.1016/j.bdr.2015.01.003

Legendre P, Fortin MJ (1989) Spatial pattern and ecological analysis. Vegetatio 80(2):107–138. https://doi.org/10.1007/BF00048036

Li S, Dragicevic S, Castro AC et al (2016) Geospatial big data handling theory and methods: a review and research challenges. ISPRS J Photogramm Remote Sens 115:119–133. https://doi.org/10.1016/j.isprsjprs.2015.10.012

Luo Q, Griffith DA, Wu H (2017) The Moran coefficient and Geary ratio: some mathematical and numerical comparisons. In: Griffith DA, Chun Y, Dean DJ (eds) Advances in geocomputation. Advances in geographic information science. Springer, Cham, pp 253–269

Moran PAP (1950) Notes on continuous stochastic phenomena. Biometrika 37(1/2):17–23. https://doi.org/10.2307/2332142

Oden D (1995) Adjusting Moran's I for population density. Stat Med 14(1):17–26

Tait M, Tobin J (2017) Three conjectures in extremal spectral graph theory. J Comb Theory Ser B 126:137–161. https://doi.org/10.1016/j.jctb.2017.04.006

Tiefelsdorf M, Boots B (1995) The exact distribution of Moran's I. Environ Plan A 27(6):985–999. https://doi.org/10.1068/a270985

van Zyl T (2014) Algorithmic design considerations for geospatial and/or temporal big data. In: Karimi HA (ed) Big data: techniques and technologies in geoinformatics. CRC Press, Baca Raton, pp 117–132

Waldhör T (1996) The spatial autocorrelation coefficient Moran's I under heteroscedasticity. Stat Med 15(7–9):887–892

Weiss NA (2017) Introductory statistics, 10th edn. Pearson Education Ltd, London

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Qing Luo[1,3] · Daniel A. Griffith[2] · Huayi Wu[1,3]

Qing Luo
luoqing11@whu.edu.cn

Daniel A. Griffith
dagriffith@utdallas.edu

[1] State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, Hubei, China

[2] School of Economic, Political, and Policy Science, The University of Texas at Dallas, Richardson, TX 75080-3021, USA

[3] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, Hubei, China