Counter-Intuitive Effect of Null Hypothesis on Moran's *I* Tests Under Heterogenous Populations

Hayato Nishi[1]
Graduate School of Social Data Science, Hitotsubashi University, Tokyo, Japan

Ikuho Yamada
Center for Spatial Information Science, The University of Tokyo, Japan

Abstract

We examine the effect of null hypothesis on spatial autocorrelation tests using Moran's I statistic. There are two possible variable states that do not exhibit spatial autocorrelation. One is that they have the same average values in all small regions, and the other is that they are not the same, but their variations are spatially random. The second state is less restrictive than the first. Thus, it intuitively appears suitable for the null hypothesis of Moran's I test. However, we found that it can make false discoveries more frequently than the nominal rate of the test when the first state is the true data generation process.

2012 ACM Subject Classification Information systems → Geographic information systems

Keywords and phrases Moran's I statistic, spatial autocorrelation, spatial heterogeneity, false discovery, null hypothesis

Digital Object Identifier 10.4230/LIPIcs.GIScience.2023.56

Category Short Paper

Funding *Hayato Nishi*: CSIS Joint Research Grants Program, Center for Spatial Information Science, the University of Tokyo.
Ikuho Yamada: JSPS KAKENHI Grant Number JP 22H00245.

1 Introduction

Moran's I statistic [3] is one of the most widely accepted statistics for testing spatial autocorrelation in spatially aggregated quantitative data such as the results of social surveys aggregated at the municipality level. A typical example of data to be tested is "per capita" quantity. For instance, we may obtain the average income of each municipality from a survey and test whether spatial clusters of high (or low) income exist using these data. In this paper, we discuss two fundamental aspects of Moran's I test that are often overlooked but can potentially affect the results of the test. One is the reliability of the observations and the other is the null hypothesis.

The reliability of the observations varies among municipalities because of their heterogeneous populations and sizes. Although the original implementation of Moran's I test does not consider such variability in data reliability, studies have pointed out its influence on results and proposed adjustment methods for heterogeneous populations [4, 7, 1].

In addition to population heterogeneity, the selection of the null hypothesis also affects the results of Moran's I test. [1] classified the spatial risk pattern (which corresponds to the income pattern in our example) to be tested into three states:

- $\mathcal{A}$. spatially constant risk,
- $\mathcal{B}$. heterogeneous risks without spatial correlation, and
- $\mathcal{C}$. heterogeneous risks with spatial correlation.

[1] Corresponding author

Although the Hypotheses $\mathcal{A}$ and $\mathcal{B}$ imply no spatial autocorrelation, their practical meanings are substantially different. Hypothesis $\mathcal{A}$ is rejected when there are differences in the average income of individual municipalities. By contrast, $\mathcal{B}$ is rejected only when the differences in average income have spatial clusters. Therefore, we consider $\mathcal{A}$ as a more rigorous state of no spatial autocorrelation than $\mathcal{B}$. When one suspects that the data in hand have a spatial pattern of $\mathcal{C}$, it appears reasonable to employ $\mathcal{B}$ as the null hypothesis to detect spatial autocorrelation in the data. Employing $\mathcal{A}$ as the null hypothesis would result in overdetection because it regards the spatial pattern of $\mathcal{B}$ as spatial autocorrelation. However, analysts do not always carefully examine the null hypothesis when applying Moran's I test. In this study, we investigate how our choice of null hypothesis and population adjustment influences the results of Moran's I test.

This paper is structured into four sections, including this introduction. Section 2 discusses the theoretical basis for adjusting Moran's I test for heterogeneous populations. Section 3 presents simulation studies using synthetic grids and population data for Japanese municipalities. Section 4 summarizes our major findings.

2 Spatial Autocorrelation Tests with Moran's *I*

2.1 Moran's *I* Statistic

Let us consider a set of observed values $\boldsymbol{x} = (x_1, \ldots, x_n)^\top$ for a study region consisting of n regions. Let $\mathbf{C}$ be a known spatial adjacency matrix and $c_{i,j}$ be its $i-j$ element. When regions i and j are adjacent, $c_{i,j} = 1$; otherwise, $c_{i,j} = 0$. Furthermore, for diagonal elements, $c_{i,i} = 0$. Let $\mathbf{W}$ be a row-standardized version of $\mathbf{C}$ and $w_{i,j}$ be the $i-j$ element. In the simulation studies discussed in Section 3, we define $\mathbf{C}$ as Queen's contiguity matrix. Using these notations, Moran's I statistic is defined as

$$I(\boldsymbol{x}) = \frac{n\boldsymbol{x}^\top \mathbf{M}\mathbf{W}\mathbf{M}\boldsymbol{x}}{W_0 \boldsymbol{x}^\top \mathbf{M}\boldsymbol{x}} \tag{1}$$

where $W_0 = \sum_i \sum_j w_{i,j}$ and $\mathbf{M} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. Note that $\mathbf{I}$ is the identity matrix of size n and $\mathbf{1}$ is an $n \times 1$ vector, all of whose elements are 1.

2.2 Data Generation Process and Null Hypothesis

Here, we derive the distribution of Moran's I when $\boldsymbol{x}$ follows the Gaussian distribution. We assume that x_i represents the estimated value of an unknown parameter μ_i. For instance, let x_i be the average income observed in region i, μ_i be its true value without biases such as measurement errors, and $y_{i,k}$ be income that an individual k in region i gains. As $y_{i,k}$ generally contains personal differences and measurement errors, we assume that $y_{i,k}$ follows a normal distribution with mean μ_i and variance σ^2. Letting m_i be the population of region i, x_i is given by $\frac{1}{m_i}\sum_k y_{i,k}$; thus, it can be discerned that the observation x_i follows a normal distribution with mean μ_i and variance $\frac{\sigma^2}{m_i}$. If the data generation process (DGP) is $\mathcal{A}$, the mean μ_i is constant μ for the entire study region. However, if DGP is $\mathcal{B}$, μ_i is not uniform. Following [1], we assume that μ_i follows an independent normal distribution of mean μ and variance $\sigma^2 s^2$. The parameter s^2 controls the relative heterogeneity of true values μ_i. If $s^2 = 0$, then the DGP corresponds to $\mathcal{A}$, whereas if $s^2 > 0$, it corresponds to $\mathcal{B}$.

Therefore, letting $\boldsymbol{\mu} = (\mu_1, ..., \mu_n)^\top$ be the vector of true income values and $\boldsymbol{\Sigma}$ be a diagonal matrix whose $i-i$ element is $\frac{1}{m_i} + s^2$, $\boldsymbol{x}$ follows a multivariate normal distribution of the mean $\boldsymbol{\mu}$ and the variance-covariance matrix $\boldsymbol{\Sigma}$. Below we explain our finding that,

when the mean μ_i is constant μ for the entire study region, and $s^2 = 0$, the distribution of Moran's I does not depend on unknown parameters μ and σ^2. When $\mathbf{\Sigma}$ can be decomposed into $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\top$ by Cholesky decomposition,

$$\boldsymbol{x} = \mu\mathbf{1} + \sigma\mathbf{L}\boldsymbol{\varepsilon} \tag{2}$$

where $\boldsymbol{\varepsilon}$ is a vector of elements following a standard normal distribution. By substituting this into x in Eq. (1), we can obtain

$$I(\boldsymbol{x}) = \frac{n\boldsymbol{x}^\top\mathbf{M}\mathbf{W}\mathbf{M}\boldsymbol{x}}{W_0\boldsymbol{x}^\top\mathbf{M}\boldsymbol{x}} = \frac{n\varepsilon^\top\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L}\varepsilon}{W_0\varepsilon^\top\mathbf{L}^\top\mathbf{M}\boldsymbol{x}} \tag{3}$$

given that $\mathbf{M}\mathbf{L} = \mathbf{0}$ and $\mathbf{M}\boldsymbol{x} = \mu\mathbf{M}\mathbf{1} + \sigma\mathbf{M}\mathbf{L}\boldsymbol{\varepsilon} = \sigma\mathbf{M}\mathbf{L}\boldsymbol{\varepsilon}$, where $\mathbf{0}$ is a zero vector. Eq. (3) includes neither the parameters μ nor σ^2, implying that Moran's I statistic is a pivotal statistic independent of the unknown parameters when we assume $\mathcal{A}$ as a null hypothesis. [5] and [6] present the distribution of Moran's I statistic and its approximation, respectively, when the observed vector $\boldsymbol{x}$ follows a normal distribution. Based on them and Eq. (3), the probability that $I(x)$ is less than an arbitrary value I_{obs} can be written as

$$\Pr\left[I(x) \leq I_{obs}\right] = \Pr\left[\boldsymbol{\varepsilon}^\top\left(n\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L} - I_{obs}W_0\mathbf{L}^\top\mathbf{M}\mathbf{L}\right)\boldsymbol{\varepsilon} \leq 0\right]. \tag{4}$$

Let $\mathbf{T}$ be $n\mathbf{L}^\top\mathbf{M}\mathbf{W}\mathbf{M}\mathbf{L} - I_{obs}W_0\mathbf{L}^\top\mathbf{M}\mathbf{L}$ and its eigenvalue decomposition be $\mathbf{T} = \mathbf{E}^\top\mathbf{\Lambda}\mathbf{E}$, where $\mathbf{\Lambda}$ is a diagonal matrix composed of the eigenvalues, $(\lambda_1, ..., \lambda_n)$. If we make $\mathbf{E}$ an orthogonal matrix, $\boldsymbol{z} = \mathbf{E}\boldsymbol{\varepsilon}$ follows independent normal distributions; thus, the left-hand side of the inequality in Eq. (4) , $\boldsymbol{\varepsilon}^\top\mathbf{C}\boldsymbol{\varepsilon} = \sum_i \lambda_i z_i^2$, follows the generalized chi-square distributions. [2] provides details of this transformation. This property indicates that we can evaluate the cumulative distribution of Moran's I statistic by evaluating that of the generalized chi-square distribution without using the unknown parameters μ and σ^2.

This property is particularly beneficial when the population m_i is not uniform because we cannot employ the permutation test approach because the observation vector $\boldsymbol{x}$ is not exchangeable. If our null hypothesis is $\mathcal{A}$, then we assume $s^2 = 0$. Hence, we can apply population adjustment without knowledge of the unknown parameters μ and σ^2. However, if our null hypothesis is $\mathcal{B}$, we need s^2 for the population adjustment.

We cannot distinguish $\mathcal{A}$ and $\mathcal{B}$ when the population m_i is uniform for the entire study region. Therefore, the selection of the null hypothesis $\mathcal{A}$ or $\mathcal{B}$ does not affect the property of Moran's I test when the populations are uniform and the DGP is Gaussian. By contrast, in the case of heterogeneous populations, it is unclear how the selection of the null hypothesis affects the results. In the next section, we examine the potential influence of this selection using two simulation studies.

3 Simulation Studies

This section describes the settings and results of the simulations. The two study regions are discussed in Sections 3.1 and 3.2. Section 3.1 presents a synthetic grid system with three population patterns to examine the influence of population heterogeneity. For a more realistic scenario, we introduce real-world municipalities and their populations in Section 3.2. Section 3.3 illustrates how the choice of null hypothesis influences the false discovery rate (FDR) of Moran's I test in our simulations.

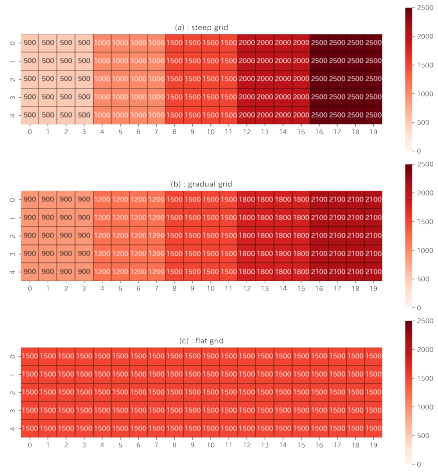

Figure 1 Populations on the Synthetic Grids.

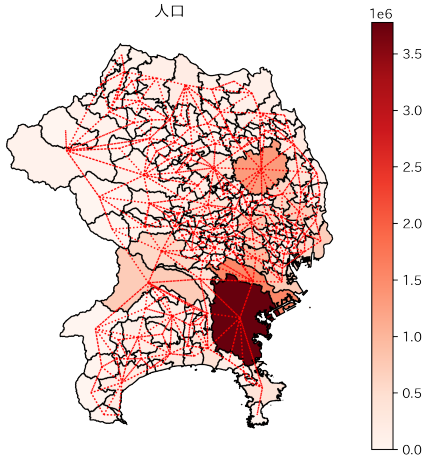

Figure 2 Municipalities and Populations in Tokyo, Japan.

3.1 Synthetic Grid Data

We consider a 20×5 regular grid system as the study region. Using the notation defined in Section 2 and assuming that an individual living in region i has the value of a target variable with mean μ_i and variance σ^2, the value of x_i can be simulated as a random number obtained from the normal distribution of mean μ_i and variance $\frac{\sigma^2}{m_i}$. Note that variance σ^2 is set constant for the entire study region. Once the local mean μ_i is marginalized, the observation x_i follows a normal distribution of the mean μ and variance $\sigma^2(\frac{1}{m_i} + s^2)$.

To examine the influence of heterogeneous populations, we consider the three spatial distributions of the regional populations shown in Figure 1. The "steep grid" pattern shown in Figure 1(a) has the regional population that steeply increases toward the right, while the "flat grid" pattern in Figure 1(c) shows a constant regional population for the entire study region. The "gradual grid" pattern in Figure 1(b) is in between; while its regional population also increases toward the right, it is less steep than the steep grid pattern. The regional populations are arranged such that the total populations are the same.

In the simulations described in Section 3.3, we set $\sigma^2 = 1.0$ and $\mu = 0$. For the nonspatial autocorrelation state $\mathcal{B}$, we select $s^2 = 1.0$.

3.2 Tokyo Municipality Data

For a realistic study region, we use municipal and population data from three prefectures in the Tokyo Metropolitan Area in Japan: Tokyo, Kanagawa, and Saitama. The municipal boundaries and populations are shown in Figure 2. The red dashed lines show the neighborhood relationships among municipalities defined by the Queen style.

In the simulations in Section 3.3, we set σ^2 as the same as the average of m_i and $\mu = 0$. For the nonspatial autocorrelation state $\mathcal{B}$, we select $s^2 = \sigma^{-2}$.

3.3 Results

For both the grids and Tokyo, we applied one-sided tests to detect a positive autocorrelation at the 5% significance level. We employed the numerical approach presented in [5] to calculate the cumulative probability that appears in Eq. (4). Therefore, numerical errors were included in the simulation results.

Table 1 False Discovery Rates on the Synthetic Grids.

(a) steep grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP=$\mathcal{A}$	0.049	0.077
DGP=$\mathcal{B}$	0.027	0.049
(b) gradual grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP=$\mathcal{A}$	0.050	0.057
DGP=$\mathcal{B}$	0.043	0.049
(c) flat grid		
	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP=$\mathcal{A}$	0.050	0.050
DGP=$\mathcal{B}$	0.050	0.050

Table 2 False Discovery Rates on Tokyo Municipalities.

	$H_0 = \mathcal{A}$	$H_0 = \mathcal{B}$
DGP=$\mathcal{A}$	0.050	0.058
DGP=$\mathcal{B}$	0.042	0.049

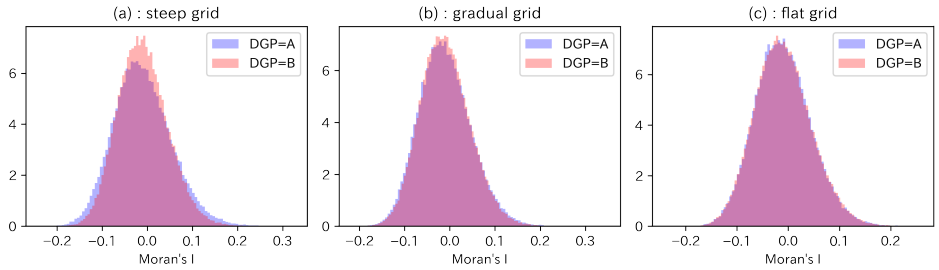

Figure 3 The Distributions of Moran's I on the Synthetic Grids.

Table 1 shows the false discovery rates (FDR) of the synthetic grids described in Section 3.1. In the case of (c) flat grid, $\mathcal{A}$ and $\mathcal{B}$ are identical, as discussed in Section 2.2. Thus, we do not need to consider differences in the null hypothesis if the population is uniform. However, in other grids, FDRs equal to a nominal rate of 5% only when the null hypothesis H_0 is correctly selected. This shows that the null hypothesis $H_0 = \mathcal{B}$, which allows heterogeneity of the true mean μ_i, results in a much higher FDR than expected, when actual μ_i is homogeneous. The opposite result is obtained when we employ $H_0 = \mathcal{A}$. Thus, counterintuitively, a test that assumes homogeneous means is more conservative than one that allows heterogeneous means. This tendency is clearer in (a) steep grid than in (b) gradual grid.

Table 2 shows the result of Tokyo municipality data. We observe the same counterintuitive results as those found in synthetic grids.

The results in Tables 1 and 2 indicate that $H_0 = \mathcal{A}$ is a safer choice than $H_0 = \mathcal{B}$ to keep FDR less than 5%, which is the predetermined nominal significance level of the test. This is because the distribution of Moran's I from $H_0 = \mathcal{A}$ exhibits a larger variance than from $H_0 = \mathcal{B}$. Figure 3 shows Moran's I distributions for the synthetic grids. However, whether this property is always observed remains unclear.

4 Conclusion

Intuitively, the test under the null hypothesis $\mathcal{B}$ does not reject it if the true data generation process (DGP) is $\mathcal{A}$. Hence, it sounds reasonable for analysts to employ $\mathcal{B}$ as their null hypothesis if they want to discover only $\mathcal{C}$. However, our simulation studies based on

synthetic grids and real municipalities with population data revealed that testing under the null hypothesis $\mathcal{B}$ does not guarantee that FDR becomes less than the nominal significance level if the true DGP is $\mathcal{A}$. In other words, if we employ $\mathcal{B}$ as a null hypothesis, we may often detect incorrect "spatial autocorrelation" of income when income is the same in all municipalities. This implies that the null hypothesis must be selected carefully when applying spatial autocorrelation test.

Further research is needed to examine whether this counterintuitive property appears in other situations, such as the target variable x_i following non-Gaussian distributions and the spatial contiguity matrix $\mathbf{C}$ different from Queen's definition. To evaluate the performance of the test, the statistical power, in addition to FDR, also needs to be examined. This is not straightforward because the true value of s^2 is generally unknown; thus, practical approaches are required.

References

1 Renato M. Assunção and Edna A Reis. A new proposal to adjust Moran's I for population density. *Statistics in Medicine*, 18(16):2147–2162, 1999.

2 Abhranil Das and Wilson S. Geisler. A method to integrate and classify normal distributions. *Journal of Vision*, 21(10):1, September 2021.

3 P. A. P. Moran. Notes on Continuous Stochastic Phenomena. *Biometrika*, 37(1/2):17, June 1950.

4 Neal Oden. Adjusting Moran's I for population density. *Statistics in Medicine*, 14(1):17–26, January 1995.

5 Michael Tiefelsdorf. Some practical applications of Moran's I's exact conditional distribution. *Papers in Regional Science*, 77(2):101–129, 1998.

6 Michael Tiefelsdorf. The saddlepoint approximation of Moran's I's and local Moran's Ii's reference distributions and their numerical evaluation. *Geographical Analysis*, 34(3):187–206, 2002.

7 Thomas Waldhör. The spatial autocorrelation coefficient Moran's I under heteroscedasticity. In *Statistics in Medicine*, volume 15, pages 887–892, 1996.