Review Paper

Easy way to understand the Global measures of Spatial Autocorrelations

Kalesh M. Karun, Amitha Puranik and Binu V.S.
Department of Statistics, Manipal University, Manipal, INDIA

Abstract

Spatial autocorrelation measures the spatial dependency of observations that quantifies the degree of spatial clustering or dispersion in the values of a variable measured across a set of locations. There are two types of spatial autocorrelation measures such as global measures and local measures. In this article we are giving an overview of Moran's I and Geary's C (Global measures) along with the steps for manual calculation and R code for the estimation of these measures.

Keywords: Spatial autocorrelation, Moran's I, Geary's C, Spatial weight matrix, Moran's Scatter plot.

Introduction

Observations from different locations or geographical areas are usually related. i.e., observations located nearby are more related than the observations located farther apart. This type of dependency of observation based on the location position is termed as spatial autocorrelation, which measures the correlation of a variable with itself through space. 'Spatial autocorrelation' is the correlation among values of a single variable strictly attributable to their relatively close positions (location) on a two-dimensional (2D) surface, introducing a deviation from the independent observations assumption of classical statistics[1]. Spatial autocorrelation coefficients indicate whether and to what extent the observations influence each other via the structure of the network.

Spatial Weight Matrix (W): Spatial autocorrelation depends on the distance between the observations in space. So to assess spatial autocorrelation, we should define distance between the observations using any of the distance functions. This distance between the observations are presented in the form of matrix, known as weight matrix, which gives the relationships between locations from where the measurements were taken. The spatial weights matrix W is an *n* x *n* nonnegative matrix that specifies the "neighborhood set" for each observation. These weights are sometimes referred to as a neighboring function.

The weight matrix can be specified in several ways[2]: i. The weight for any two different locations is a constant. ii. All observations within a specified distance have fixed weight. iii. K nearest neighbors has a fixed weight, and all others are zero. iv. Weight is proportional to inverse distance, inverse distance squared, or inverse distance up to a specified distance. v. Euclidean, Manhattan, and Minkowski distance are common distance functions in spatial analysis[3] and these distance can be calculated by the formula

$$d_{ij} = \left[(p-k)^m + (q-l)^m \right]^{1/m}$$

Where, *d* is the distance between i^{th} and j^{th} spatial observations, (*p, q*) is the co-ordinate position of ith observation, (*k, l*) is the co-ordinate position of jth observation.

The generic 'm' parameter in the above equation can be replaced by the value 2 to obtain Euclidean distance; 1 to obtain the Manhattan distance, and all the intermediate values between 1 and 2 to obtain array of Minkowski distances. Euclidean distance is most often used distance function in spatial analysis and inverse of the distance is taken as weight which is given as,

$$W_{ij} = 1/d_{ij} = \frac{1}{\left[(p-k)^2 + (q-l)^2 \right]^{1/2}}$$

Example - Estimation of weight matrix: Consider a spatial arrangement of 5 districts of a city as shown below. Here the districts 2 and 3 are neighbors to districts 1and 4, whereas district 5 has only one neighbor, district 4.
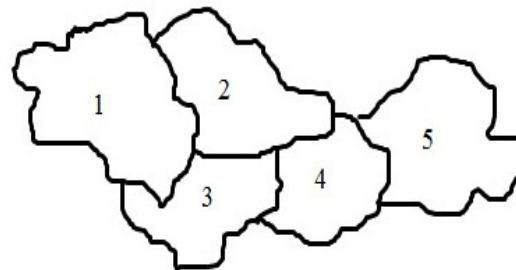

Figure–1
Spatial arrangement of five districts of a city

A convenient way to summarize the spatial configuration of 5 districts is the weight matrix, which is given below,

$$W = \begin{bmatrix} 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

In its simplest form, these weights will take values 1 for close neighbors, and 0 otherwise. By convention, the diagonal elements of the weights matrix are set to zero[4], ie $w_{ii} = 0$.

Weight matrix based on the Euclidian distance to summarize the spatial configuration of 5 districts is given as,

$$W = \begin{bmatrix} 0 & 1 & 0.5 & 0.33 & 0.25 \\ 1 & 0 & 1 & 0.5 & 0.33 \\ 0.5 & 1 & 0 & 1 & 0.5 \\ 0.33 & 0.5 & 1 & 0 & 1 \\ 0.25 & 0.33 & 0.5 & 1 & 0 \end{bmatrix}$$

Measures of Global spatial autocorrelation measures

There are two indicators of spatial autocorrelation measure such as global measures and local measures. Global spatial autocorrelation is a measure of the overall clustering of the data which provides one correlation statistic to summarize the whole study area. But if there is no global autocorrelation or no clustering in the whole area, one can look for clusters at a local level using the measure known as local spatial autocorrelation. Moran's I and Geary's C is the measures of global spatial autocorrelation[5].

Moran's I: Moran's I tests for global spatial autocorrelation for continuous data. It is based on the cross-products of deviations of observations from their mean and is calculated by accounting for location of the observations[6]. Moran's I formula is related to Pearson's correlation coefficient, which is a measure of correlation between two variables. Moran's I deals with the correlation of values of a single variable. In Pearson's r, the denominator is the product of the standard deviations of the two variables, whereas in Moran's I there is only one variable involved.

Moran's I formula is:

$$I = \frac{n}{S_0} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (x_i - \overline{x})(x_j - \overline{x}) w_{ij}}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

where $\overline{x}$ is the mean of the variable x, n is the number of observations in the variable x, w_{ij} is the distance based weight given to each pair *(i, j)*, and is given by

$$w_{ij} = \frac{1}{\sqrt{(p-k)^2 + (q-l)^2}}$$ where *(p, q)* is the co-ordinate position of ith observation and *(k,l)* is the co-ordinate position of jth observation, S_0 is the sum of the weights and is given by

$$S_0 = \sum_i \sum_j w_{ij}$$

We can rewrite the equation as,

$$I = \frac{n}{s_0} \frac{X^T W X}{X^T X}$$

where,

$$X = \begin{bmatrix} x_1 - \overline{x} & x_2 - \overline{x} & . & . & . & x_n - \overline{x} \end{bmatrix}$$

and

$$W = \begin{bmatrix} w_{11} & w_{12} & . & . & . & w_{1n} \\ w_{21} & w_{22} & . & . & . & w_{2n} \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ w_{n1} & w_{n2} & . & . & . & w_{nn} \end{bmatrix}$$

Range of Moran's I: Moran's I usually takes values in the interval [–1, +1], although values lower than –1 or higher than +1 may occasionally be obtained. Positive spatial autocorrelation occurs when similar values occur near one another. Negative spatial autocorrelation occurs when dissimilar values occur near one another[3] and approximately zero when the observations are independently arranged in space.

The extreme values of Moran's I depends on the maximum and minimum eigen values of the weight matrix. If the weight matrix *W* is symmetric, then the extreme values of the quadratic form $X^T W X / X^T X$ for all *X* are simply the smallest and largest eigen value of W^7. If the weight matrix W is not symmetric, the extreme values of the quadratic form can be found by noting that $X^T W X = X^T S X$, where $S = (W + W^T)/2$, so the extreme values of the quadratic form is the smallest and largest eigen values of *S*. Usually the minimum and maximum of Moran's I ranges from -1 to $1^{6,8}$.

$I_{min} = (n/S_0)*$ Smallest eigenv alue of W
$I_{max} = (n/S_0)*$ Largest eigen value of W

Example - Calculation of Moran's I statistic: If 3, 5, 4 and 2 are the values measured for a variable in (1, 1), (1, 2), (2, 1) and (2, 2) co-ordinate positions as shown below,

3	5
4	2

Solution

Table–1
Calculation of Moran's I

x_i	x_j	$x_i - \bar{x}$	$x_j - \bar{x}$	w_{ij}	$(x_i - \bar{x})(x_j - \bar{x})\,w_{ij}$
3	3	-0.5	-0.5	0	0
3	5	-0.5	1.5	1	-0.75
3	4	-0.5	0.5	1	-0.25
3	2	-0.5	-1.5	0.71	0.5325
5	3	1.5	-0.5	1	-0.75
5	5	1.5	1.5	0	0
5	4	1.5	0.5	0.71	0.5325
5	2	1.5	-1.5	1	-2.25
4	3	0.5	-0.5	1	-0.25
4	5	0.5	1.5	0.71	0.5325
4	4	0.5	0.5	0	0
4	2	0.5	-1.5	1	-0.75
2	3	-1.5	-0.5	0.71	0.5325
2	5	-1.5	1.5	1	-2.25
2	4	-1.5	0.5	1	-0.75
2	2	-1.5	-1.5	0	0
Sum				10.84	-5.87

Here S_0= 10.84 and N=4

$$\sum_{i=1}^{4}(x_i - \bar{x})^2 = (3-3.5)^2 + (5-3.5)^2 + (4-3.5)^2 + (2-3.5)^2 = 5$$

I= 4* -5.87 / 10.84 * 5 = **-0.4332**

Table–2
R code for the Moran's I

```
x=c(3,5,4,2)   # observations
W=matrix(c(0,1,1,.71,1,0,.71,1,1,.71,0,1,.71,1,1,0),length
(x)) #Weight matrix
X=x-mean(x)
Xt=t(X)
I=(length(X)*Xt%*%W%*%X)/(sum(W)*Xt%*%X)
I #Estimate of Moran's I
```

Moran scatter plot: The Moran scatter plot[7] is a visual tool for exploratory analysis. It enables to assess how related an observed value is to its neighboring observations. Its horizontal axis is based on the values of the observations and is also known as the response axis. The vertical axis is based on the weighted average or spatial lag of the corresponding observations.

Steps to construct the Moran scatter plot: i. Calculate standardized scores (Z_x) based on the spatial data values (i.e., subtract each observation from the mean and then divide by the standard deviation) ii. Construct the weight matrix (*W*). iii. Calculate the spatial lag values (Wx_i) (i.e., multiplying the weight matrix by the spatial data values). iv. Calculate the standardized spatial lag scores (Z_y) (i.e., subtract each spatial lag value from the mean and then divide by the standard

deviation). v. Construct the scatter plot by placing Z_x in the horizontal axis and Z_y in the vertical axis

Example – Moran's scatter plot: If 3, 5, 4 and 2 are the values measured for a variable in (1, 1), (1, 2), (2, 1) and (2, 2) co-ordinate positions, then the Moran's scatter plot is given as below,

3	5
4	2

The weight matrix is given by $W = \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$

Table–3
Calculation of standardized scores

Observations x_i	Standardized scores (Z_x) $(x_i - \bar{x})/\sigma$	Spatial lag (Wx_i)	Standardized spatial lag scores (Z_y) $(W_{x_i} - \overline{W}_x)/\sigma_{W_{xi}}$
3	-0.3873	9	0.8696
5	1.1618	5	-0.8696
4	0.3873	5	-0.8696
2	-1.1618	9	0.8696

Here $\bar{x} = 3.5$; $\sigma = 1.29$; $\overline{W}_x = 7$; $\sigma_{W_{x_i}} = 2.3$

The scatter plot obtained by plotting the Standardized scores (Z_x) in the horizontal axis and the Standardized spatial lag score (Z_y) in the vertical axis gives the Moran scatter plot as below.
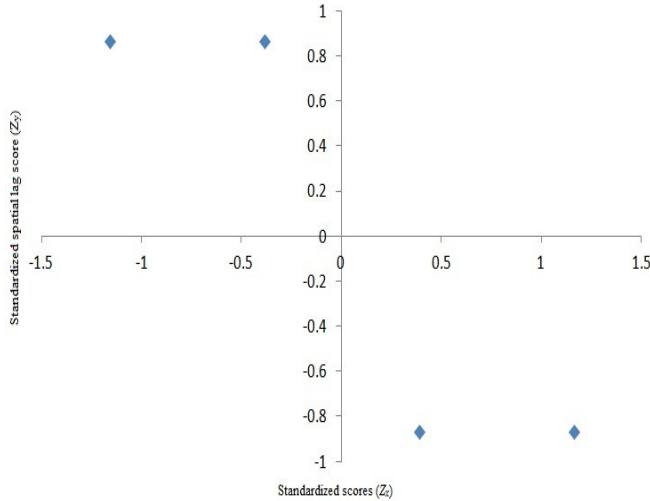
Figure–2
Moran's scatter plot

The four quadrants in the graph provide a classification of four types of spatial autocorrelation: high-high (upper right), low-low (lower left), or positive spatial autocorrelation; high-low (lower right) and low-high (upper left), for negative spatial autocorrelation.

Geary's C: Geary's C is second popular indicator of global spatial autocorrelation. Geary's C statistic (Geary 1954) is based on the deviations in responses of each observation with one another[6, 8] and is given as,

$$ C = \frac{n-1}{2S_0} \frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \overline{x})^2} $$

where, *n* is the number of observations for the variable *x*, $\overline{x}$ is the mean of the variable *x*, w_{ij} is the distance based weight given to each pair *(i, j)*, and is given by $w_{ij} = \dfrac{1}{\sqrt{(p-k)^2 + (q-l)^2}}$ where *(p, q)* is the co-ordinate position of ith observation and *(k,l)* is the co-ordinate position of jth observation, S_0 is the sum of the weights and is given by $S_0 = \sum_i \sum_j w_{ij}$

Range of Geary's C: The value of Geary's *C* lies between 0 and 2. Values lower than 1 demonstrates increasing positive spatial autocorrelation, values higher than 1 indicate increasing negative spatial autocorrelation and 1 means no spatial autocorrelation. Its expectation is 1 in the absence of autocorrelation and regardless of the specified weight matrix[8].

The extreme values of Geary's C depend on the maximum and minimum eigen values of the weight matrix.

I_{min} = (n-1 /2S_0)* Smallest eigen value of B
I_{max} = (n-1 /2S_0)* Largest eigen value of B
where $B_{ij}=(R_i+K_j)\delta_{ij}-2W_{ij}$
(δ_{ij} is the kronecker delta (δ_{ij} =0 if i ≠j and =1 if i=j) , R_i is row sum of the ith row of W, K_j is the column sum of the jth column of W)

Example - Calculation of Geary's C: If 3, 5, 4 and 2 are the values measured for a variable in (1, 1), (1, 2), (2, 1) and (2, 2) co-ordinate positions, then one can calculate the correlation between these observations as,

3	5
4	2

Table–4
Calculation of Geary's C

x_i	x_j	x_i-x_j	$(x_i$-$x_j)^2$	w_{ij}	$(x_i$-$x_j)^2*w_{ij}$
3	3	0	0	0	0
3	5	-2	4	1	4
3	4	-1	1	1	1
3	2	1	1	0.71	0.71
5	3	2	4	1	4
5	5	0	0	0	0
5	4	1	1	0.71	0.71
5	2	3	9	1	9
4	3	1	1	1	1
4	5	-1	1	0.71	0.71
4	4	0	0	0	0
4	2	2	4	1	4
Sum				**10.84**	**38.84**

Here S_0= 10.84 and N=4

$$ \sum_{i=1}^{4}(x_i - \overline{x})^2 = (3-3.5)^2 + (5-3.5)^2 + (4-3.5)^2 + (2-3.5)^2 = 5 $$

$$ C = \frac{3*38.84}{2*10.84*5} = 1.0749 $$

Table–5
R code for the Geary's C statistic

```
x=c(3,5,4,2)  # Observations
W=matrix(c(0,1,1,.71,1,0,.71,1,1,.71,0,1,.71,1,1,0),4)  # Weight matrix
y=matrix((rep(x,4)),length(x))
X=x-mean(x)
C=(length(x)-1)*sum(W*((y-t(y))^2))/(2*sum(W)*t(X)%*%X)
C  # Estimate of Geary's Statistic
```

Conclusion

Statistics relies on observations being independent from one another. If observations are clustered or dependent on each other, the assumption of independence is violated and in such

situations the spatial autocorrelation measures are appropriate. Morans I and Geary's C are the two important measures of global spatial autocorrelation. Moran's I is sensitive to extreme values, whereas Geary's C is more sensitive to differences in small neighborhoods. Even though Moran's I and Geary's C result in similar conclusions, Moran's I is preferred in most cases since it is consistently more powerful than Geary's C[5].

References

1. Griffith D.A., Spatial structure and spatial interaction: 25 years later, *The Review of Regional Studies.,* **37(1),** 28-38 **(2007)**

2. Nilupa G., Liu Y. and Park J., Spatial autocorrelation, Purdue University Department of Statistics (online)

3. Rizwan S., Stefania B., Merril L.K. and William A.G., Comparison of distance measures in spatial analytical modeling for health service planning, *BMC health services research*., **9(1),** 200 **(2009)**

4. Luc A. and Sergio J.R., Perspectives on spatial data analysis, Springer, **(2010)**

5. Cliff A.D. and Ord J.K., Model building and the analysis of spatial pattern in human geography, *Journal of the Royal Statistical Society*. Series B (Methodological)., 297–348 **(1975)**

6. Bellman R.E., Introduction to matrix analysis, SIAM., 960, **(1970)**

7. Anselin L., An introduction to spatial autocorrelation analysis with GeoDa. Spatial Analysis Laboratory, University of Illinois, Champagne-Urbana, Illinois **(2003)**

8. Jong P.D., Sprenger C. and Veen F., On extreme values of Moran's I and Geary's c, *Geographical Analysis.*, **16(1),** 17-24 **(1984)**