Measuring Spatial Autocorrelation with GRASS
A Tutorial

James Darrell McCauley, USDA Fellow
Department of Agricultural Engineering
Purdue University

This document describes the basics of spatial autocorrelation indices (Geary Ratio and Moran Coefficient) and the GRASS commands to do a simple spatial autocorrelation analysis. All of the equations and much of the verbage are from *Spatial Autocorrelation—A Primer* by Daniel A. Griffith (1981), which is published by the Association of American Geographers (tel: 202.234.1450).
GRASS commands: `s.voronoi`, `v.autocorr`

1 Spatial Autocorrelation

Spatial autocorrelation refers to the spatial ordering of a single variable and to the relationship between pairs of observations of this variable. The ordering of n observed values of some variable X is usually described with the aid of a connectivity matrix, $\mathbf{C}$. Non–zero c_{ij} entries in the $n \times n$ matrix indicate that the corresponding row and column units (polygons) are juxtaposed. For example, consider the arrangement in Figure 1.

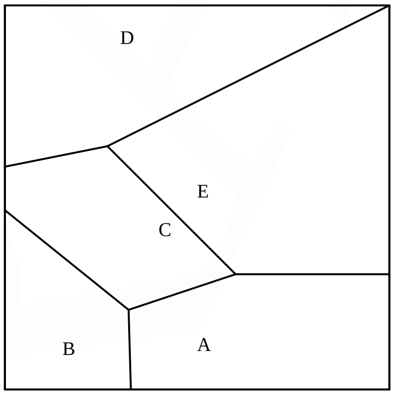


Figure 1: A Simple Map

Polygon A shares line segments with polygons B, C, and E. Hence, c_{ij} entries associate with these pairs would be non–zero. Since polygon A does not share a line segment with D, the matrix entry associate with this pair of observations would be zero. The connectivity matrix for the above arrangement is:

$$\mathbf{C} = \begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{pmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array}$$

Note that, in general, there are relationships other than sharing a line segment which may give a non–zero entry, though this is the most common. Also, non–zero entries do not have to be unity. They could be used to express more meaning than just connectivity (*e.g.*, transport costs).

1.1 The Randomization Assumption

Consider the possibility of enumerating all possible ways that the values in the previous map could be arranged. Mathematically, it could be expressed as

$$\frac{n!}{\prod\limits_{j=1}^{n}(n_j!)} \tag{1}$$

and would represent the population from which spatial autocorrelation statistics are computed. Usually, only the statistics associated with the *observed* map pattern are of concern.

1.2 The Normality Assumption

Instead of assuming that the observed map pattern is a single sample from the set of all possible permutations, attributes for each polygon may be treated as a random sample from a population within that polygon (or area unit). In other words, the map would be considered a joint realization of n random variables X_i.

1.3 Spatial Autocorrelation Indices

If data are measured on a nominal scale (*i.e.*, binary, *has* or *doesn't have* a characteristic), then the *join count* statistic is used. Much more interesting, though, are data measured on an interval/ratio scale. For this type of data, other statistics are applied.

1.3.1 The Geary Ratio

The Geary Ratio is an index for interval/ratio data that is based upon paired comparisons of juxtaposed map values. It may be calculated as

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

Assuming a randomization perspective, if this index is calculated for all $n!\left(\prod_{j=1}^{n}(n_j!)\right)^{-1}$ arrangements, the average of these values would be one (*i.e.*, $\mathrm{E}[GR] = 1$). For the normalization assumption, the expected value is also one.

The meaning is this index is fairly straighforward: as similar values tend to clump together ($c_{ij} = 1$), then the Geary Ratio approaches zero ($(x_i - x_j)^2 \to 0$). If dissimiliar x_i values tend to clump together, the Geary Ratio approaches 2 (and sometimes more).

1.3.2 The Moran Coefficient

Another index for interval/ratio data is the Moran Coefficient and may be calculated as

$$MC = \frac{n\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} c_{ij}\,(x_i - \bar{x})(x_j - \bar{x})}{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{n} c_{ij}\sum\limits_{i=1}^{n}(x_i - \bar{x})^2}. \tag{3}$$

The expected value of the MC is $\frac{-1}{n-1}$. For similar values in juxtapostion, $MC \rightarrow 1$. For dissimiliar values in juxtaposition, $MC \rightarrow -1$. As the Geary Ratio deals with paired comparisons, the Moran Coefficient deals with covariations.

1.3.3 Sampling Distributions

In order to estimate population means of the Geary Ratio and the Moran Coefficient, Monte Carlo simulation methods are normally used rather than enumerating the population of statistics. Under the randomization assumption, the population is sampled by using random permutations of the set of $\{x_i\}$ values. Normally, this is done until the sampled values of GR and MC conform very closely to a normal distribution (tested by using a Wilk–Shapiro statistic or some other appropriate test for normality).

Under the normality assumption, the mean (μ) and standard deviation (σ) of X are estimated from the observed values. Then, sets of $\{x_i\}$ are created using deviates from $N\left(\mu, \sigma^2\right)$. New sets are created until the sampled values of GR and MC conform to a normal distribution.

The standard errors of the Geary Ratio and the Moran Coefficient can be calculated directly (*i.e.,* from observed values). Under the normality assumption, the equations are

$$\sigma^2_{GR} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} + \frac{1}{2}\sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\left(\sum_{j=1}^{n} c_{ij} - 1\right)\right)(n-1) - \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2}{(n+1)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}/2\right)^2} \tag{4}$$

and

$$\begin{aligned}\sigma^2_{MC} &= \frac{2n^2\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} - 4n\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} + \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\left(\sum_{j=1}^{n} c_{ij} - 1\right)\right)\right) + 3\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2}{\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2 (n^2-1)} \\ &\quad - \left(\frac{-1}{n-1}\right)^2 .\end{aligned} \tag{5}$$

Under the randomization assumption, these become more complex:

$$\sigma^2_{GR} = \frac{\left[\frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2\left(-(n-1)^2\frac{m_4}{s^4} + \left(n^2-3\right)\right) + \sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}(n-1)\left(-(n-1)\frac{m_4}{s^4} + n^2 - 3n + 3\right) + \frac{1}{2}\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} + \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\left(\sum_{j=1}^{n} c_{ij} - 1\right)\right)\right)(n-1)\left(\left(n^2-n+2\right)\frac{m_4}{s^4} - \left(n^2+3n+6\right)\right)\right]}{\frac{1}{2}n(n-2)(n-3)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2/2} \tag{6}$$

where

$$\begin{aligned} m_4 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4 \\ s^4 &= \left(\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2\right)^2 , \end{aligned}$$

and

$$\sigma^2_{MC} = \frac{\left[2n\left(\left(n^2-3n+3\right)\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right) - 4n^2\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} + \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\left(\sum_{j=1}^{n} c_{ij}-1\right)\right)\right) + 3n\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2 - \frac{m_4}{s^4}\left(2\left(n^2-n\right)\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} - 8n\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij} + \sum_{i=1}^{n}\left(\sum_{j=1}^{n} c_{ij}\left(\sum_{j=1}^{n} c_{ij}-1\right)\right)\right) + 6\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2\right)\right]}{\left(n-1\right)\left(n-2\right)\left(n-3\right)\left(\sum_{i=1}^{n}\sum_{j=1}^{n} c_{ij}\right)^2} - \left(\frac{-1}{n-1}\right)^2. \quad (7)$$

1.4 Testing for Spatial Autocorrelation

Suppose that we want to test whether or not there is non–zero spatial autocorrelation in the population. Because the standard error formula do not use any sample statistics, a z test can be used. Recall the form of a standard normal:

$$z = \left(GR - \mu_{GR}\right)/\sigma_{GR} = \left(GR-1\right)/\sigma_{GR} \quad (8)$$

$$z = \left(MC - \mu_{MC}\right)/\sigma_{MC} = \left(MC + 1/\left(n-1\right)\right)/\sigma_{MC} \quad (9)$$

The pair of null and alternative hypotheses are:

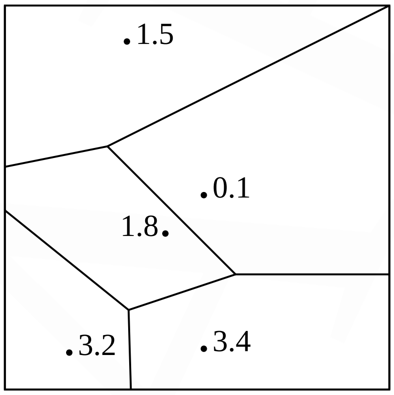
$$H_O : \mu_{MC} = -1/\left(n-1\right) \qquad H_O : \mu_{GR} = 1$$
$$H_A : \mu_{MC} \neq -1/\left(n-1\right) \qquad H_A : \mu_{GR} \neq 1$$

2 GRASS Utilities

Often a difficult stumbling block in analysis of spatial autocorrelation is the first step: creation of a connectivity matrix. Given only punctual data, it is first necessary to create Theissen or Voronoi polygons before connectivity is expressed. Fortunately, there is a utility for GRASS 4.0 which does this called `s.voronoi`.[1] Sample input and output is shown in figure 2.[2]

```
name|tutorial
desc|sample points
10|50|#1 3.4
09|15|#2 3.2
40|40|#3 1.8
90|30|#4 1.5
50|50|#5 0.1
```

(a) sites file



(b) binary vector file

Figure 2: Labeled vector map created from sites list.

The creation of a connectivity matrix, as well as calculation of spatial autocorrelation indices, distributions parameters, and z statistics for calculating H_O, can be handled by the GRASS utility `v.autocorr`. Using the map in figure 2 as input, the following output was obtained:

[1] Available via anonymous ftp from `moon.cecer.army.mil`

[2] Figure drawn by adding PostScript commands to `v.out.ascii` output.

```
SPATIAL AUTOCORRELATION REPORT
        Location: [ttt]
        Mapset: [PERMANENT]
        n=5 labeled polygons in [tutorial]

        Geary Ratio=0.798043           Moran Coefficient=-0.00684932

Sampling Distributions under different assumptions
        Randomization: Mean(GR)=1.00459        Mean(MC)=-0.252668
                       Var(GR)=0.0435003       Var(MC)=0.0353123
        Normalization: Mean(GR)=0.993255       Mean(MC)=-0.249718
                       Var(GR)=0.047619        Var(MC)=0.0327381

Significance Tests
        Ho: non-zero spatial autocorrelation is present
        Ho: Mean(MC) = -1(n-1)         Ho: Mean(GR) = 1
        Ha: Mean(MC) != -1(n-1)        Ha: Mean(GR) != 1

        Randomization: Z(MC)=(-0.00684932+0.25)/0.187916
                       Z(MC)=1.29394
        Randomization: Z(GR)=(0.798043-1)/0.208567
                       Z(GR)=-0.968306
        Normalization: Z(MC)=(-0.00684932+0.25)/0.180937
                       Z(MC)=1.34384
        Normalization: Z(GR)=(0.798043-1)/0.218218
                       Z(GR)=-0.925483
```

For this small example, we cannot reject H_O at a 10% significance level ($z = \pm 1.645$).

3 Summary

Spatial autocorrelation indicates if similar or dissimilar valued units tend to be juxtaposed. The Geary Ratio and Moran Coefficient are two common indices of spatial autocorrelation for interval/ratio data. The ability to calculate these indices and estimate their means and standard deviations has been added to GRASS, as well as the ability to create Voronoi diagrams as binary vector files from sites lists.

The commands used to perform the analysis in this short tutorial are:

```
s.voronoi site=tutorial vect=tutorial    # create labeled vector file from sites
v.support tutorial                       # create support files
v.autocorr tutorial                      # spatial autocorrelation analysis
```

The connectivity matrix given at the beginning of this document was obtained by specifying the `-cn` flags to `v.autocorr`. Source code for these new GRASS commands may be obtained via anonymous ftp from `moon.cecer.army.mil`.

It should be pointed out that, at present, the number of samples used to estimate μ_{MC}andμ_{GR} (during Monte Carlo simulation) are constant. A more appropriate approach to this task would be to continue sampling until a Wilk–Shapiro (or other test statistic) indicates that the distribution closely approximates normal. This addition is planned for the future, as well as calculation of critical z values for specified significance levels, which would give the software the ability to indicate whether or not H_O can be rejected.

Please forward any comments regarding this tutorial or these GRASS commands to

`mccauley@ecn.purdue.edu` or `mccauley%ecn@purccvm`.

Also, please consult others works on spatial statistics if you plan to use these commands.

Acknowledgements

The author is grateful for financial support provided by Purdue University and the United States Department of Agriculture, for the many people on the GIS–L mailing list, a.k.a. `comp.infosystems.gis`, for initially educating me about spatial autocorrelation, and to Daniel Griffith, whose writing encouraged me to add these features to GRASS.