Applied Spatial Statistics in R, Section 2

Spatial Autocorrelation

Yuri M. Zhukov

IQSS, Harvard University

January 16, 2010

Outline

What is Spatial Autocorrelation?

- Spatial autocorrelation measures the degree to which a phenomenon of interest is correlated to itself in space (Cliff and Ord 1973, 1981).
- Tests of spatial autocorrelation examine whether the observed value of a variable at one location is independent of values of that variable at neighboring locations.
- Positive spatial autocorrelation indicates that similar values appear close to each other, or cluster, in space
- Negative spatial autocorrelation indicates that neighboring values are dissimilar or, equivalenty, that similar values are dispersed.
- Null spatial autocorrelation indicates that the spatial pattern is random.

Global autocorrelation: Moran's $\mathcal{I}$

- The Moran's $\mathcal{I}$ coefficient calculates the ratio between the product of the variable of interest and its spatial lag, with the product of the variable of interest, adjusted for the spatial weights used.

$$\mathcal{I} = \frac{n}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}(y_i - \bar{y})(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from –1 (perfect dispersion) to $+1$ (perfect correlation). A zero value indicates a random spatial pattern.
- Under the null hypothesis of no autocorrelation, $\mathbb{E}[\mathcal{I}] = \frac{-1}{n-1}$

Global autocorrelation: Moran's $\mathcal{I}$

- Calculating the variance of Moran's $\mathcal{I}$ is a little more involved:

$$Var(\mathcal{I}) = \frac{n\mathfrak{s}_1 - \mathfrak{s}_2\mathfrak{s}_3}{(n-1)(n-2)(n-3)(\sum_i \sum_j w_{ij})^2}$$

$$\mathfrak{s}_1 = (n^2 - 3n + 3)\Big(\frac{1}{2}\sum_i \sum_j (w_{ij} + w_{ji})^2\Big) - n\Big(\sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2\Big) + 3(\sum_i \sum_j w_{ij})^2$$

$$\mathfrak{s}_2 = \frac{n^{-1}\sum_i (y_i - \bar{x})^4}{(n^{-1}\sum_i (y_i - \bar{x})^2)^2}$$

$$\mathfrak{s}_3 = \frac{1}{2}\sum_i \sum_j (w_{ij} + w_{ji})^2 - 2n\Big(\frac{1}{2}\sum_i \sum_j (w_{ij} + w_{ji})^2\Big) + 6\Big(\sum_i \sum_j w_{ij}\Big)^2$$

Global autocorrelation: Geary's $\mathcal{C}$

- The Geary's $\mathcal{C}$ uses the sum of squared differences between pairs of data values as its measure of covariation.

$$\mathcal{C} = \frac{(n-1)\sum_i \sum_j w_{ij}(y_i - y_j)^2}{2(\sum_i \sum_j w_{ij})\sum_i (y_i - \bar{y})^2}$$

- where y_i is the value of a variable for the ith observation, $\bar{y}$ is the sample mean and w_{ij} is the spatial weight of the connection between i and j.
- Values range from 0 (perfect correlation) to 2 (perfect dispersion). A value of 1 indicates a random spatial pattern.

Global autocorrelation: Join Counts

- When the variable of interest is *categorical*, a join count analysis can be used to assess the degree of clustering or dispersion.
- A binary variable is mapped in two colors (Black & White), such that a join, or edge, is classified as either *WW* (0-0), *BB* (1-1), or *BW* (1-0).
- Join count statistics can show
  - positive spatial autocorrelation (clustering) if the number of *BW* joins is significantly *lower* than what we would expect by chance,
  - negative spatial autocorrelation (dispersion) if the number of *BW* joins is significantly *higher* than what we would expect by chance,
  - null spatial autocorrelation (random pattern) if the number of *BW* joins is approximately *the same* as what we would expect by chance.

Global autocorrelation: Join Counts

- By the naive definition of probability, if we have n_B Black units and $n_W = n - n_B$ White units, the respective probabilities of observing the two types of units are:

$$P_B = \frac{n_B}{n} \qquad P_W = \frac{n - n_B}{n} = 1 - P_B$$

- The probabilities of BB and WW in two adjacent cells are

$$P_{BB} = P_B P_B = P_B^2 \qquad P_{WW} = (1 - P_B)(1 - P_B) = (1 - P_B)^2$$

- The probability of BW in two adjacent cells is

$$P_{BW} = P_B(1 - P_B) + (1 - P_B)P_B = 2P_B(1 - P_B)$$

Global autocorrelation: Join Counts

- The <u>expected counts</u> of each type of join are:

$$\mathbb{E}[BB] = \frac{1}{2}\sum_i \sum_j w_{ij} P_B^2 \qquad \mathbb{E}[WW] = \frac{1}{2}\sum_i \sum_j w_{ij}(1 - P_B)^2$$

$$\mathbb{E}[BW] = \frac{1}{2}\sum_i \sum_j w_{ij} 2P_B(1 - P_B)$$

- Where $\frac{1}{2}\sum_i \sum_j w_{ij}$ is the total number of joins (of any type) on a map, assuming a binary connectivity matrix.
- The <u>observed counts</u> are:

$$BB = \frac{1}{2}\sum_i \sum_j w_{ij} y_i y_j \qquad WW = \frac{1}{2}\sum_i \sum_j w_{ij}(1 - y_i)(1 - y_j)$$

$$BW = \frac{1}{2}\sum_i \sum_j w_{ij}(y_i - y_j)^2$$

- where $y_i = 1$ if unit i is Black and $y_i = 0$ if White.

Global autocorrelation: Join Counts

- The variance of BW is calculated as

$$
\begin{aligned}
\sigma^2_{BW} =& \mathbb{E}[BW^2] - \mathbb{E}[BW]^2 \\
=& \frac{1}{4}\left(\frac{2\mathfrak{s}_2 n_B(n-n_B)}{n(n-1)} + \frac{(\mathfrak{s}_3 - \mathfrak{s}_1) n_B(n-n_B)}{n(n-1)} \right. \\
& \left. + \frac{4(\mathfrak{s}_1^2 + \mathfrak{s}_2 - \mathfrak{s}_3) n_B(n_B-1)(n-n_B)(n-n_B-1)}{n(n-1)(n-2)(n-3)} \right) - \mathbb{E}[BW]^2 \\
\mathfrak{s}_1 =& \sum_i \sum_j w_{ij} \\
\mathfrak{s}_2 =& \frac{1}{2} \sum_i \sum_j (w_{ij} - w_{ji})^2 \\
\mathfrak{s}_3 =& \sum_i (\sum_j w_{ij} + \sum_j w_{ji})^2
\end{aligned}
$$

Global autocorrelation: Join Counts

- A test statistic for the BW join count is

$$\mathcal{Z}(BW) = \frac{BW - \mathbb{E}[BW]}{\sqrt{\sigma^2_{BW}}}$$

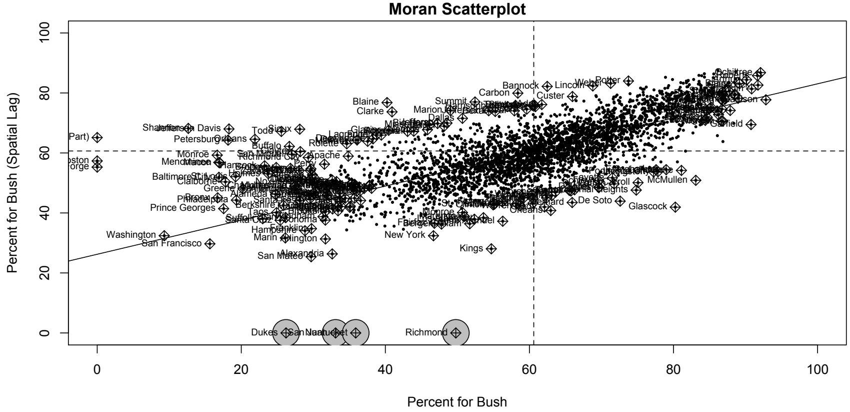
- The join count statistic is assumed to be asymptotically normally distributed under the null hypothesis of no spatial autocorrelation.
- The test of significance is then provided by evaluating the BW statistic as a standard deviate (Cliff and Ord, 1981).

Local autocorrelation

- Global tests for spatial autocorrelation are calculated from local relationships between observed values at spatial units and their neighbors.
- It is possible to break these measures down into their components, thus constructing local tests for spatial autocorrelation.
- These tests can be used to detect
  - Clusters, or units with similar neighbors
  - Hotspots, or units with dissimilar neighbors

Local autocorrelation

Below is a scatterplot of county vote for Bush and its spatial lag (average vote received in neighboring counties). The Moran's $\mathcal{I}$ coefficient is drawn as the slope of the linear relationship between the two. The plot is partitioned into four quadrants: low-low, low-high, high-low and high-high.



Local autocorrelation: Local Moran's $\mathcal{I}$

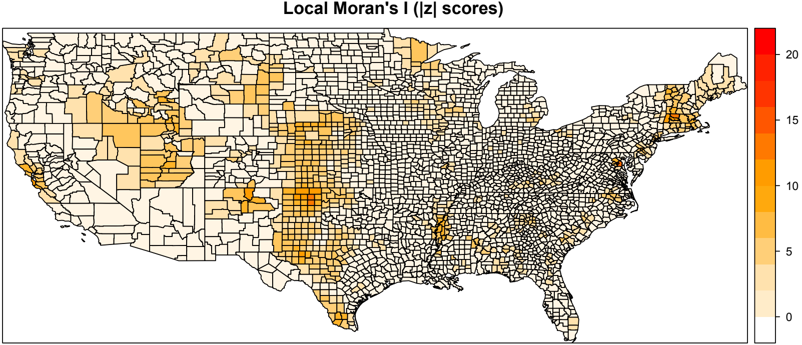
- A local Moran's $\mathcal{I}$ coefficient for unit i can be constructed as one of the n components which comprise the global test:

$$\mathcal{I}_i = \frac{(y_i - \bar{y}) \sum_{j=1}^{n} w_{ij}(y_j - \bar{y})}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}}$$

- As with global statistics, we assume that the global mean $\bar{y}$ is an adequate representation of the variable of interest.
- As before, local statistics can be tested for divergence from expected values, under assumptions of normality.

Local autocorrelation: Local Moran's $\mathcal{I}$

Below is a plot of Local Moran $|z|$-scores for the 2004 Presidential Elections. Higher absolute values of z scores (red) indicate the presence of "hot spots", where the percentage of the vote received by President Bush was significantly different from that in neighboring counties.



Words of Caution

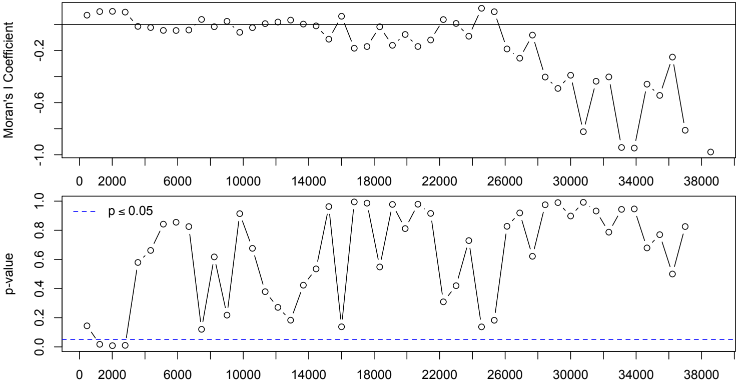
1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.

Words of Caution

1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.

Words of Caution

Below is a correlogram of Moran's $\mathcal{I}$ coefficients for Polity IV country democracy scores in 2008. The x-axis represents distances between country capitals, in kilometers. Here, democracy is significantly ($p \leq .05$) spatially autocorrelated only at distances of 3,000 km and below. So, autocorrelation estimates will depend highly on choice of lag distance.



Words of Caution

1. Autocorrelation tests are highly sensitive to spatial patterning in the variable of interest from any source. But by assuming that the mean model removes such systematic spatial patterning, spatial autocorrelation tests do not always produce useful insights into the DGP.
2. These tests are also highly sensitive to one's choice of spatial weights. Where the weights do not reflect the "true" structure of spatial interaction, estimated autocorrelation (or lack thereof) may actually stem from misspecification.
3. As originally designed, spatial autocorrelation tests assumed there are no neighborless units in the study area. When this assumption is violated, the size of n may be adjusted (reduced) to reflect the fact that some units are effectively being ignored. Not doing so will generally bias the absolute value for the autocorrelation statistic upward and the variance downward.

Examples in R

Switch to R tutorial script. Section 2.