Environmental Statistics

Slides - Part 2

Fedele Greco

Department of Statistical Sciences
University of Bologna

Email: fedele.greco@unibo.it

Contents

1. Testing for spatial autocorrelation

2. Local Indicators of Spatial Association (LISA)

Environmental Statistics - Part 2

1. Testing for spatial autocorrelation

2. Local Indicators of Spatial Association (LISA)

Testing for spatial autocorrelation

We will discuss hypothesis testing using Moran's *I* as the test statistic.

Moran's *I* is the spatial counterpart of the Durbin-Watson test statistic in time series analysis for testing temporal autocorrelation.

Burridge (1980) has shown that the test based on Moran's *I* is asymptotically equivalent to the likelihood ratio test.

Testing for spatial autocorrelation means testing the following hypotheses system:

$$\begin{cases} H_0 : \rho = 0 \\ H_1 : \rho > 0 \end{cases}$$

Moran's I - Moments under H_0

It can be shown that:

$$E(I|H_0) = -\frac{1}{n-1}$$

Under the hypothesis that the study variable Y follows a Normal distribution:

$$V_N(I|H_0) = E(I^2|H0) - [E(I|H0)]^2 = \frac{nS_1 - nS_2 + 3S_0^2}{(n-1)(n-1)S_0^2} - [E(I|H0)]^2$$

If the normality assumption is dropped:

$$V_R(I|H_0) = \frac{n\left((n^2-3n+3)S_1 - nS_2 + 3S_0^2\right) - K(S_1 - nS_2 + 3S_0^2)}{(n-1)(n-1)S_0^2}$$

Randomisation variance: is the variance that one would obtain by computing I on all the $n!$ permutations of the data.

Moran's I - Moments under H_0

The quantities S_0, S_1 and S_2 depends on $\boldsymbol{W}$

$$S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$$

$$S_1 = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_{i=1}^{n} (w_{i+} + w_{+i})^2$$

K is the kurtosis of $\boldsymbol{y}$

$$K = n \frac{\sum_{i=1}^{n} (y_i - \bar{y})^4}{\left(\sum_{i=1}^{n} (y_i - \bar{y})^2\right)^2}$$

Moran's I - Distribution under H_0

Note that $V(I|H_0)$ does not depend on y under the normality assumption.

If normality of the study variable is assumed, the distribution of the standardized test statistic is *asymptotically* Standard Normal, i.e.:

$$z = \frac{I - E(I|H_0)}{\sqrt{V_N(I|H_0)}} \sim N(0, 1)$$

This can be used as a z-score to test H_0.

If the study variable is not Gaussian (and in particular if its distribution is heavily skewed) the asymptotic approximation to the distribution of the standardized test statistic can be very poor.

Permutation test

Statistical significance of the spatial autocorrelation can be tested without any assumption on the distribution of the test statistic by means of permutation tests.

This can be achieved by comparing the observed value of the index with the (n!) values obtained by data permutation. The strategy proceeds as follows:

1. data are permuted over the study region
2. for each permutation, the value of the index is computed
3. a p-value can be computed as the proportion of values obtained by permutation that exceed the observed value

Monte Carlo Permutation test

Computing the value of the index under all possible permutations is impractical, particularly when n is large.

As an example, considering the Italian provinces, $110! = 1.588246e + 178$ (i.e. a permutation test based on all possible permutations is just not feasible).

The rationale of Monte Carlo permutation tests is to draw a sample of K permutations from all possible permutations and to use the $K << n!$ samples to compute the p-value.

Monte Carlo Permutation test

Let I^{*k} be the values of the test statistic computed at the k-th permutation, $k = 1, ..., K$. The p-value is approximated as:

$$Pr(I > I^{obs}|H_0) = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}(I^{*k} > I^{obs})$$

where $\mathbb{I}(I^{*k} > I^{obs}) = 1$ if $I^{*k} > I^{obs}$ and $\mathbb{I}(I^{*k} > I^{obs}) = 0$ otherwise

Environmental Statistics - Part 2

1 Testing for spatial autocorrelation

2 Local Indicators of Spatial Association (LISA)

Local Indicators of Spatial Association (LISA)

We will discuss Local indicators of spatial association (LISA) as proposed in Anselin, (1995).

LISA allow for the decomposition of global indicators, such as Moran's *I*, into the contribution of each area.

This class of indicators is useful because it allows:

- the assessment of significant local spatial clustering around an individual location;
- the indication of pockets of spatial non-stationarity, or the suggestion of spatial outliers.

Local Indicators of Spatial Association (LISA)

As an operational definition, Anselin suggested that a local indicator of spatial association is any statistic that satisfies the following two requirements:

1. the LISA for each observation gives an indication of the extent of significant spatial clustering of similar values around that observation;
2. the sum of LISAs for all observations is proportional to a global indicator of spatial association.

Formally, a local indicator is built as:

$$L_i = f(y_i, y_j, j \in N(i))$$

where f is a function to be specified. Note that, local indicators are function of the values observed at the i-th area and at its neighbors only.

Local Indicators of Spatial Association (LISA)

The L_i should be such that it is possible to infer the statistical significance of the pattern of spatial association at location i.

The second requirement of a LISA, that is, its relation to a global statistic, is stated formally as:

$$\sum_{i=1}^{n} L_i = \gamma \Lambda \tag{1}$$

where Λ is a global indicator of spatial association and γ is a scale factor.

By means of LISA, **local spatial clusters** as well as **spatial outliers**, may be identified as those locations or sets of contiguous locations for which the LISA is significant.

The general LISA can be used as the basis for a test on the null hypothesis of no local spatial association, although general results on the distribution of a generic LISA are hard to obtain.

This is similar to the problems encountered in deriving distributions for global statistics, for which typically only approximate or asymptotic results are available. An alternative is the use of a conditional permutation approach to yield Monte Carlo p-values.

LISA: the local Moran's I

The local Moran's I is defined as follows:

$$I_i = \frac{(y_i - \bar{y}) \sum_{j=1}^{n} \widetilde{w}_{ij}(y_j - \bar{y})}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \quad i = 1, \ldots, n$$

Note that:

$$I = \sum_{i=1}^{n} I_i$$

i.e. local Moran's I is coherent with requirement (1).

LISA: the local Moran's I

It turns out that:

$I_i > 0$ if $(y_i - \bar{y})$ has the same sign of $\sum_{j=1}^{n} \widetilde{w}_{ij}(y_j - \bar{y}) = (\widetilde{y}_i - \bar{y})$,

while

$I_i < 0$ if $(y_i - \bar{y})$ has opposite sign with respect to $\sum_{j=1}^{n} \widetilde{w}_{ij}(y_j - \bar{y}) = (\widetilde{y}_i - \bar{y})$,

LISA: the local Moran's *I*

The local Moran's *I* is strictly related to the Moran scatter plot. The quadrants of the Moran scatter plot allow to classify the areas according to different type of spatial patterns.

	Q.	Corr.	Interpretation
high-high	I	+	I'm high and my neighbors are high
high-low	II	-	I'm a high outlier among low neighbors
low-low	III	+	I'm low and my neighbors are low
low-high	IV	-	I'm a low outlier among high neighbors

where Q denotes the Quadrant.

Testing for local spatial correlation

In the context of local analysis, it make sense to test the following hypothesis system:

$$\begin{cases} H_0 : \rho_i = 0 \\ H_1 : \rho_i \neq 0 \end{cases}, \quad i = 1, \ldots, n.$$

The alternative hypothesis is two-sided because, at a local level, both positive and negative spatial association are of interest in applications.

We will use local Moran's *I* as the test statistic.
The `localmoran` function of the `spdep` package works under the normality assumption for the distribution of the test statistic.

Testing for local spatial corr. - Permutation approach

The distribution of I_i is actually unknown and the Gaussian approximation is rather questionable.

According to Anselin (1995), an alternative is the use of a conditional randomization or permutation approach to yield empirical so-called pseudo significance levels.
The randomization is conditional in the sense that the value y_i at a location i is held fixed and the remaining values are randomly permuted over the locations in the data set.

For each of these resampled data sets, the value of I_i can be computed. The resulting empirical distribution function provides the basis for a statement about the extremeness (or lack of extremeness) of the observed statistic, relative to the values computed under the null hypothesis (the randomly permuted values).

Multiple testing - A very quick discussion

A complicating factor in the assessment of significance of LISAs is that the statistics for individual locations will tend to be spatially correlated.

In general, whenever the neighborhood sets $N(i)$ and $N(j)$ of two locations i and j contain common elements, the corresponding L_i and L_j will be correlated. Due to this correlation, and the associated problem of multiple comparisons, the usual interpretation of significance will be flawed.

Indeed, since we are testing simultaneously n hypotheses, the problem of **multiple testing** arises.

Multiple testing - A very quick discussion

When the overall significance associated with the multiple comparisons (correlated tests) is set to α, and there are n comparisons, the individual significance α_i should be set by taking into account the number of comparisons.

By fixing $\alpha_i = \alpha$, the number of expected rejections under the null hypothesis is αn.
If the tests are independent and the Type-I error probability of each test is set to $\alpha_i = \alpha$, the probability to reject at least one hypothesis is:

$$\begin{aligned} Pr(\text{at least 1 rejection}) &= 1 - Pr(\text{no rejections}) \\ &= 1 - (1 - \alpha)^n \end{aligned}$$

If $n = 110$ and $\alpha = .05$, $1 - (1 - \alpha)^n = 0.99$.

Multiple testing - A very quick discussion

The first method for tackling the multiple comparisons problem was proposed by Bonferroni:

$$\alpha_{iB} = \frac{\alpha}{n}$$

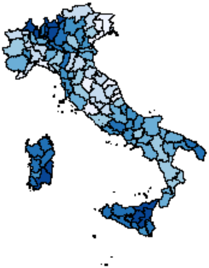
With this correction, if $n = 110$ and $\alpha = .05$, $1 - (1 - \alpha_B)^{110} = 0.048 \approxeq \alpha$. The Bonferroni correction is a conservative correction, in the sense that the correction comes at the cost of increasing the probability of producing false negatives, i.e., reducing statistical power.

An alternative method is known as the Sidak method:

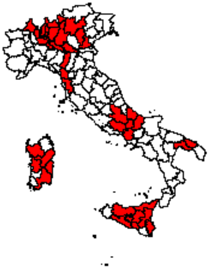
$$\alpha_{iS} = 1 - (1 - \alpha)^{\frac{1}{n}}.$$

Separate Waste Collection - Local Spatial Analysis

Under the normality assumption for the teststatistic.



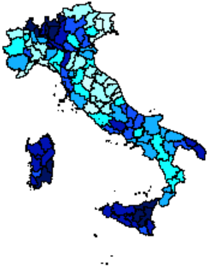
Local Moran



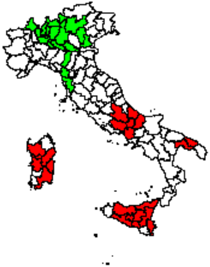
p-value<.05

Separate Waste Collection - Local Spatial Analysis

A more intuitive representation.



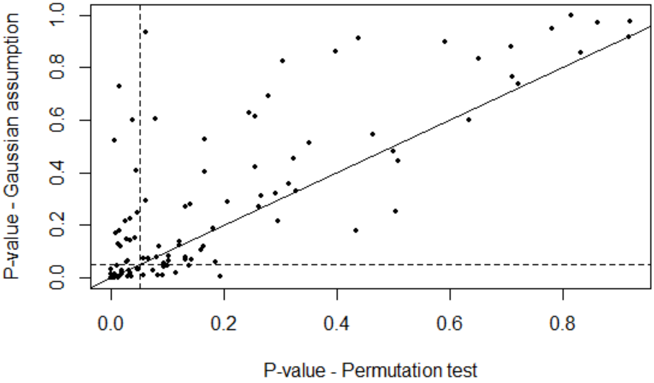
Local Moran



p-value<.05

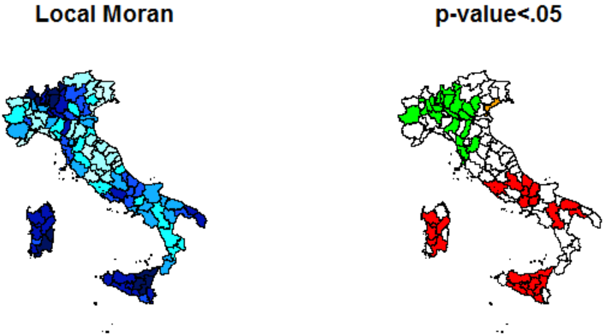
Separate Waste Collection - Local Spatial Analysis

Comparison between p-values obtained with the Gaussian assumption and with the permutation test.



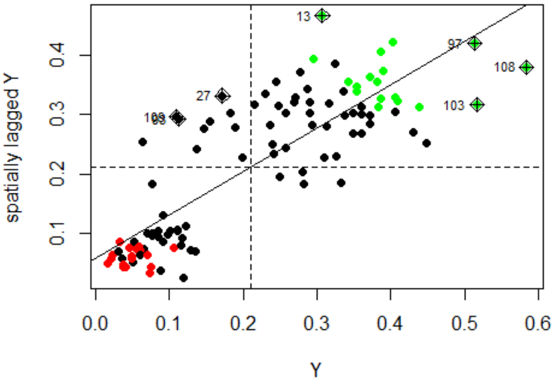
Separate Waste Collection - Local Spatial Analysis

Significant p-values obtained with the permutation test.



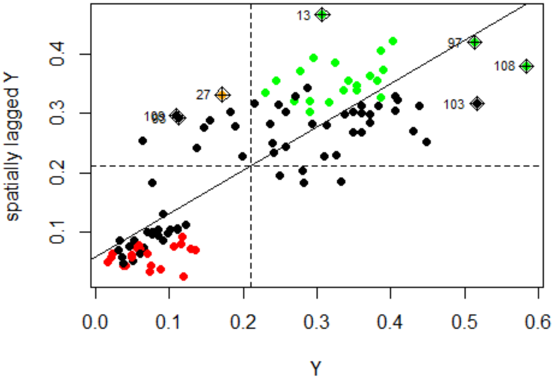
Separate Waste Collection - Local Spatial Analysis

Moran scatter plot. Areas with significant p-values (with the Gaussian assumption) are highlighted in green (high-high) and red (low-low).



Separate Waste Collection - Local Spatial Analysis

Moran scatter plot. Areas with significant p-values (according to the permutation test) are highlighted in green (high-high) and red (low-low).


spatially lagged Y
Y