Lecture 1: Introduction to Spatial Econometric

Mauricio Sarrias

Universidad de Talca

Version: 2025

Goals

At the end of this lecture, students are expected to be able to

- Understand the concepts of spatial heterogeneity and spatial autocorrelation.
- Understand the concept of the Spatial Weight Matrix $\boldsymbol{W}$.
- Construct spatial weight matrices in **R**.
- Derive and understand the main test for spatial autocorrelation.
- Perform the Moran's I test in R.

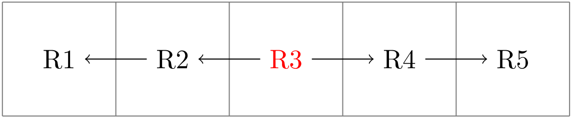
Reading for: Introduction to Spatial Econometrics

- Chapter 1 of class notes.
- Dall'Erba, S. (2005). Distribution of regional income and regional funds in Europe 1989-1999: an exploratory spatial data analysis.The Annals of Regional Science, 39(1), 121-148.
- Celebioglu, F., & Dall'erba, S. (2010). Spatial disparities across the regions of Turkey: an exploratory spatial data analysis. The Annals of Regional Science, 45(2), 379-400.

Why do We Need Spatial Econometric?

- Important aspect when studying spatial units (cities, regions, countries)
  - Potential relationships and interactions between them.
- Example: Modeling pollution:
  - Should we analyze regions as independent units?
  - No, regions are spatially interrelated by ecological and economic interactions.
  - Existence of environmental externalities:
    - an increase in i's pollution will affect the pollution in neighbors regions, but the impact will be lower for more distance regions.

Figure: Environmental Externalities



Distance Matters

Key Point:

First law of geography of Waldo Tobler: *"everything is related to everything else", but near things are more related than distant things.*

This first law is the foundation of the fundamental concepts of **spatial dependence** and **spatial autocorrelation**.

Figure: Professor Waldo Tobler

Why do We Need Spatial Econometric?

- Spatial econometric deals with spatial effects
  1. Spatial heterogeneity

Definition (Spatial heterogeneity)

Spatial heterogeneity relates to a differentiation of the effects of space over the sample units. Formally, for spatial unit i:

$$y_i = f(x_i)_i + \epsilon_i \implies y_i = \beta_i x_i + \epsilon_i$$

Lack of stability over the geographical space.

Why do We Need Spatial Econometric?

- Spatial econometric deals with spatial effects

  1. Spatial dependence

Definition (Spatial dependence)

What happens in i depends on what happens in j. Formally,

$$y_i = f(y_i, y_j) + \epsilon_i, \forall i \neq j.$$

Spatial Dependence

How would you model this situation?

Figure: Environmental Externalities

$$\text{R1} \longleftarrow \text{R2} \longleftarrow {\color{red}\text{R3}} \longrightarrow \text{R4} \longrightarrow \text{R5}$$

Spatial Dependence

Using our previous example, we would like to estimate

$$\begin{aligned}
y_1 &= \beta_{21}y_2 + \beta_{31}y_3 + \beta_{41}y_4 + \beta_{51}y_5 + \epsilon_1 \\
y_2 &= \beta_{12}y_1 + \beta_{32}y_3 + \beta_{42}y_4 + \beta_{52}y_5 + \epsilon_2 \\
y_3 &= \beta_{13}y_1 + \beta_{23}y_2 + \beta_{43}y_4 + \beta_{53}y_5 + \epsilon_3 \\
y_4 &= \beta_{14}y_1 + \beta_{24}y_2 + \beta_{34}y_3 + \beta_{54}y_5 + \epsilon_4 \\
y_5 &= \beta_{15}y_1 + \beta_{25}y_2 + \beta_{35}y_3 + \beta_{45}y_5 + \epsilon_4
\end{aligned} \tag{1}$$

where β_{ji} is the effect of pollution of region j on region i.

Spatial Dependence

Using our previous example, we would like to estimate

$$
\begin{aligned}
y_1 &= \beta_{21} y_2 + \beta_{31} y_3 + \beta_{41} y_4 + \beta_{51} y_5 + \epsilon_1 \\
y_2 &= \beta_{12} y_1 + \beta_{32} y_3 + \beta_{42} y_4 + \beta_{52} y_5 + \epsilon_2 \\
y_3 &= \beta_{13} y_1 + \beta_{23} y_2 + \beta_{43} y_4 + \beta_{53} y_5 + \epsilon_3 \\
y_4 &= \beta_{14} y_1 + \beta_{24} y_2 + \beta_{34} y_3 + \beta_{54} y_5 + \epsilon_4 \\
y_5 &= \beta_{15} y_1 + \beta_{25} y_2 + \beta_{35} y_3 + \beta_{45} y_5 + \epsilon_4
\end{aligned}
\tag{1}
$$

where β_{ji} is the effect of pollution of region j on region i.

What is the problem with this modeling strategy?

Spatial Dependence

Under standard econometric modeling, it is impossible to model spatial dependency.

Spatial Autocorrelation

- Autocorrelation $\implies$ the correlation of a variables with itself

Spatial Autocorrelation

- Autocorrelation $\implies$ the correlation of a variables with itself
  - Time series: the values of a variable at time t depends on the value of the same variable at time $t-1$.

Spatial Autocorrelation

- Autocorrelation $\implies$ the correlation of a variables with itself
  - Time series: the values of a variable at time t depends on the value of the same variable at time $t-1$.
  - Space: the correlation between the value of the variable at two different locations.

Spatial Autocorrelation

- Autocorrelation $\implies$ the correlation of a variables with itself
  - Time series: the values of a variable at time t depends on the value of the same variable at time $t - 1$.
  - Space: the correlation between the value of the variable at two different locations.

Definition (Spatial Autocorrelation)

- Correlation between the same attribute at two (or more) different locations.
- Coincidence of values similarity with location similarity.
- Under spatial dependency it is not possible to change the location of the values of certain variable without affecting the information in the sample.
- It can be positive and negative.

Spatial Autocorrelation

Definition (Positive Autocorrelation)

Observations with high (or low) values of a variable tend to be clustered in space.

Spatial Autocorrelation

Definition (Positive Autocorrelation)

Observations with high (or low) values of a variable tend to be clustered in space.

Figure: Positive Autocorrelation

1	1	
1	1	

Spatial Autocorrelation

Definition (Negative Autocorrelation)

Locations tend to be surrounded by neighbors having very dissimilar values.

Spatial Autocorrelation

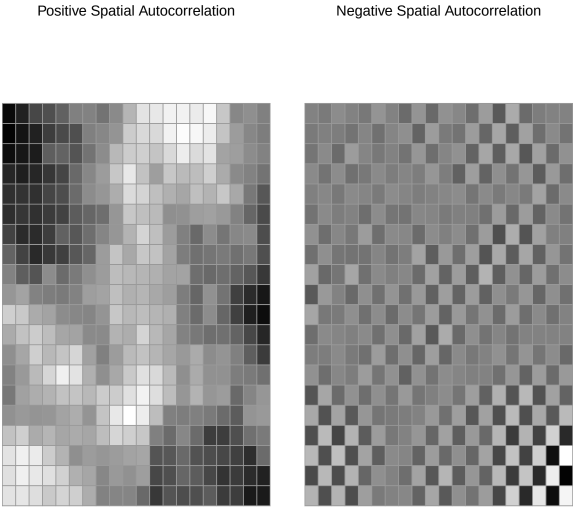
Definition (Negative Autocorrelation)

Locations tend to be surrounded by neighbors having very dissimilar values.

Figure: Negative Autocorrelation

1		1
	1	
1		1

Spatial Autocorrelation: Another Example



Spatial Autocorrelation

Definition (Spatial Randomness)

When none of the two situations occurs.

Spatial Autocorrelation

Two main sources of spatial autocorrelation (Anselin, 1988):

- Measurement errors.
- Importance of Space.

The second source is of much more interest.

Figure: Professor Luc Anselin

Why the space matters?

- The essence of regional sciences and new economic geography is that location and distance matter.
- What is observed at one point is determined by what happen elsewhere in the system.

First Law of Geography Again

Tobler's First Law of Geography

Everything depends on everything else, but closer things more so

- Important ideas:

First Law of Geography Again

Tobler's First Law of Geography

Everything depends on everything else, but closer things more so

- Important ideas:
  - Existence of Spatial Dependence.

First Law of Geography Again

Tobler's First Law of Geography

Everything depends on everything else, but closer things more so

- Important ideas:
  - Existence of Spatial Dependence.
  - Structure of Spatial Dependence

First Law of Geography Again

Tobler's First Law of Geography

Everything depends on everything else, but closer things more so

- Important ideas:
  - Existence of Spatial Dependence.
  - Structure of Spatial Dependence
    - Distance decay.

First Law of Geography Again

Tobler's First Law of Geography

Everything depends on everything else, but closer things more so

- Important ideas:
  - Existence of Spatial Dependence.
  - Structure of Spatial Dependence
    - Distance decay.
    - Closeness = Similarities.

Spatial Weight Matrix

One crucial issue in spatial econometric is the problem of formally incorporating spatial dependence into the model.

Spatial Weight Matrix

One crucial issue in spatial econometric is the problem of formally incorporating spatial dependence into the model.

Question?

What would be a good criteria to define closeness in space? Or, in other words, how to determine which other units in the system influence the one under consideration?

Spatial Weight Matrix

- The device typically used in spatial analysis is the so-called spatial weight matrix, or simply $\boldsymbol{W}$ matrix.
- Impose a structure in terms of what are the neighbors for each location.
- Assigns weights that measure the intensity of the relationship among pair of spatial units.
- Not necessarily symmetric.

Spatial Weight Matrix

Definition ($\boldsymbol{W}$ Matrix)

Let n be the number of spatial units. The spatial weight matrix, $\boldsymbol{W}$, a $n \times n$ positive symmetric and **non-stochastic** matrix with element w_{ij} at location i, j. The values of w_{ij} or the weights for each pair of locations are assigned by some preset rules which defines the spatial relations among locations. By convention, $w_{ij} = 0$ for the diagonal elements.

The symmetry assumption can be dropped later.

$$\boldsymbol{W} = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ w_{21} & w_{22} & \dots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \dots & w_{nn} \end{pmatrix}$$

Spatial Weight Matrix

Two main approaches:

1. Contiguity.
2. Based on distance

Weights Based on Boundaries

The availability of polygon or lattice data permits the construction of contiguity-based spatial weight matrices. A typical specification of the contiguity relationship in the spatial weight matrix is

$$w_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are contiguous} \\ 0 & \text{if } i \text{ and } j \text{ are not contiguous} \end{cases} \tag{2}$$

1. Binary Contiguity:
   - Rook criterion (Common Border)
   - Bishop criterion (Common Vertex)
   - Queen criterion (Either common border or vertex)

Rook Contiguity

How are the neighbors of region 5?

Figure: Rook Contiguity

1	2	3
4	5	6
7	8	9

Rook Contiguity

Figure: Rook Contiguity

1	2	3
4	5	6
7	8	9

Rook Contiguity

Figure: Rook Contiguity

1	2	3
4	5	6
7	8	9

Common border: 2, 4, 5, 6

Bishop Contiguity

Figure: Bishop Contiguity

1	2	3
4	5	6
7	8	9

Bishop Contiguity

Figure: Bishop Contiguity

1	2	3
4	5	6
7	8	9

Common vertex: 1, 3, 7, 9

Queen Contiguity

Figure: Queen Contiguity

1	2	3
4	5	6
7	8	9

Queen Contiguity

Figure: Queen Contiguity

1	2	3
4	5	6
7	8	9

Common vertex and border: 1, 2, 3, 4, 6, 7, 8, 9.

Rook Contiguity

1	2	3
4	5	6
7	8	9

$$\boldsymbol{W} = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ \color{red}{0} & \color{red}{1} & \color{red}{0} & \color{red}{1} & \color{red}{0} & \color{red}{1} & \color{red}{0} & \color{red}{1} & \color{red}{0} \\ 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \end{pmatrix}$$

Bishop Contiguity

1	2	3
4	5	6
7	8	9

$$\boldsymbol{W} = \begin{pmatrix} 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

W based on distance

- Weights may be also defined as a function of the distance between region i and j, d_{ij}.
- d_{ij} is usually computed as the distance between their centroids (or other important unit).
- Let x_i an x_j be the longitud and y_i and y_j the latitude coordinates for region i and j, respectively:

Distance Metric

Definition (Minkowski metric)

Let two point i and j, with respective coordinates (x_i, y_i) and (x_j, y_j):

$$d_{ij}^{p} = \left(|x_i - x_j|^p + |y_i - y_j|^p\right)^{1/p} \tag{3}$$

Definition (Euclidean metric)

Consider Minkowski metric and set $p = 2$, then

$$d_{ij}^{e} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}. \tag{4}$$

Definition (Manhattan metric)

Consider Minkowski metric and set $p = 1$, then

$$d_{ij}^{m} = |x_i - x_j| + |y_i - y_j|. \tag{5}$$

Distance Metric

- Euclidean distance is not necessarily the shortest distance if you take into account the curvature of the earth.

Definition (Great Circle Distance)

Let two point i and j, with respective coordinates (x_i, y_i) and (x_j, y_j):

$$d_{ij}^{cd} = r \times \arccos^{-1}\left[\cos|x_i - x_j|\cos y_i \cos y_j + \sin y_i \sin y_j\right] \tag{6}$$

where r is the Earth's radius. The arc distance is obtained in miles with $r = 3959$ and in kilometers with $r = 6371$.

$\boldsymbol{W}$ based on distance

- Inverse distance:

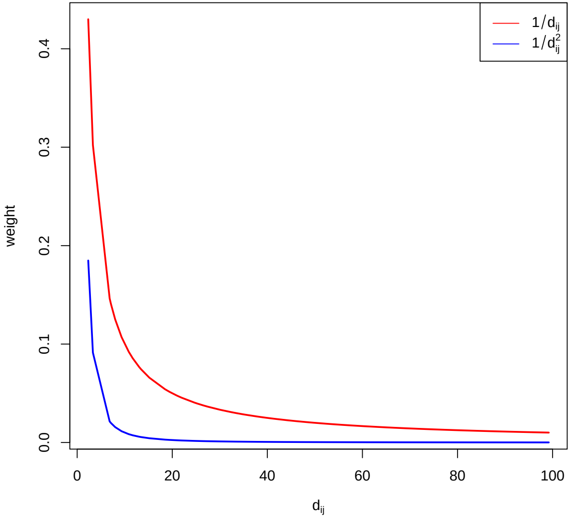
$$w_{ij} = \begin{cases} \frac{1}{d_{ij}^{\alpha}} & \text{if } i \neq j \\ 0 & \text{if } i = j \end{cases} \tag{7}$$

  Typically, $\alpha = 1$ or $\alpha = 2$.

- Negative exponential model:

$$w_{ij} = \exp\left(-\frac{d_{ij}}{\alpha}\right) \tag{8}$$

W based on distance



W based on distance

- k-nearest neighbors: We explicitly limit the number of neighbors.

$$w_{ij} = \begin{cases} 1 & \text{if centroid of } j \text{ is one of the } k \text{ nearest centroids to that of } i \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

- Threshold Distance (Distance Band Weights): In contrast to the k-nearest neighbors method, the threshold distance specifies that an region i is neighbor of j if the distance between them is less than a specified maximum distance:

$$w_{ij} = \begin{cases} 1 & \text{if } \ 0 \leq d_{ij} \leq d_{max} \\ 0 & \text{if } \ d_{ij} > d_{max} \end{cases} \tag{10}$$

To avoid isolates that would result from too stringent a critical distance, the distance must be chosen such that each location has at least one neighbor. Such a distance conforms to a max-min criterion, i.e., it is the largest of the nearest neighbor distances.

Row standardization

- $\boldsymbol{W}$'s are used to compute weighted averages in which more weight is placed on nearby observations than on distant observations.
- The elements of a row-standardized weights matrix equal

  $$w_{ij}^{s} = \frac{w_{ij}}{\sum_j w_{ij}}.$$

  This ensures that all weights are between 0 and 1 and facilities the interpretation of operation with the weights matrix as an averaging of neighboring values.
- Under row-standardization, the element of each row sum to unity.
- The row-standardized weights matrix also ensures that the spatial parameter in many spatial stochastic processes are comparable between models.
- Under row-standardization the matrices are not longer symmetric!.

Row standardization

The row-standardized matrix is also known in the literature as the row-stochastic matrix:

Definition (Row-stochastic Matrix)

A real $n \times n$ matrix $\boldsymbol{A}$ is called **Markov** matrix, or **row-stochastic matrix** if

1. $a_{ij} \geq 0$ for $1 \leq i, j \leq n$;
2. $\sum_{j=1}^{n} a_{ij} = 1$ for $1 \leq i \leq n$

An important characteristic of the row-stochastic matrix is related to its eigen values:

Theorem (Eigenvalues of row-stochastic Matrix)

Every eigenvalue ω_i of a row-stochastic Matrix satisfies $|\boldsymbol{\omega}| \leq 1$

Therefore, the eigenvalues of the row-stochastic (i.e., row-normalized, row standardized or Markov) neighborhood matrix $\boldsymbol{W}^s$ are in the range $[-1, +1]$.

Spatial Lag

The spatial lag operator takes the form $\boldsymbol{y}_L = \boldsymbol{W}\boldsymbol{y}$ with dimension $n \times 1$, where each element is given by $\boldsymbol{y}_{Li} = \sum_j w_{ij} y_j$, i.e., a weighted average of the $\boldsymbol{y}$ values in the neighbor of i.
For example:

$$\boldsymbol{W}\boldsymbol{y} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 50 \\ 30 \end{pmatrix} = \begin{pmatrix} 50 \\ 10 + 30 \\ 50 \end{pmatrix} \tag{11}$$

Using a row-standardized weight matrix:

$$\boldsymbol{W}^s\boldsymbol{y} = \begin{pmatrix} 0 & 1 & 0 \\ 0.5 & 0 & 0.5 \\ 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} 10 \\ 50 \\ 30 \end{pmatrix} = \begin{pmatrix} 50 \\ 5 + 15 \\ 50 \end{pmatrix} \tag{12}$$

Therefore, when $\boldsymbol{W}$ is standardized, each element $(\boldsymbol{W}^s\boldsymbol{y})_i$ is interpreted as a weighted average of the y values for i's neighbors.

Higher-Order Neighbors

- How to define higher-order neighbors?
  - We might be interested in the neighbors of the neighbors of spatial unit i.
- We define the higher-order spatial weigh matrix l as $\boldsymbol{W}^l$.
  - Spatial weight of order $l = 2$ is given by $\boldsymbol{W}^2 = \boldsymbol{W}\boldsymbol{W}$.
  - Spatial weight of order $l = 3$ is given by $\boldsymbol{W}^3 = \boldsymbol{W}\boldsymbol{W}\boldsymbol{W}$.
- As an illustration consider the following structure for our previous example:

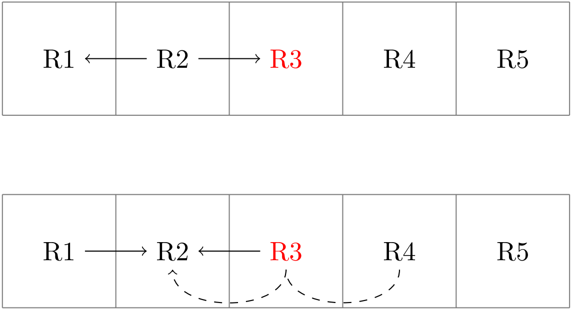
$$\boldsymbol{W} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \end{pmatrix} \tag{13}$$

Higher-Order Neighbors

Then $\boldsymbol{W}^2 = \boldsymbol{W}\boldsymbol{W}$ based on the 5×5 first-order contiguity matrix $\boldsymbol{W}$ from (13) is:

$$\boldsymbol{W}^2 = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 2 & 0 & 1 & 0 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 1 & 0 & 2 & 0 \\ 0 & 0 & 1 & 0 & 1 \end{pmatrix} \tag{14}$$

Figure: Higher-Order Neighbors



Global Autocorrelation

- Indicators of spatial association
  1. Global Autocorrelation
  2. Local Autocorrelation

Definition (Global Autocorrelation)

It is a measure of overall clustering in the data. It yields only one statistic to summarize the whole study area (Homogeneity).

1. Moran's I.
2. Gery's C.
3. Getis and Ord's $G(d)$

Definition (Local Autocorrelation)

A measure of spatial autocorrelation for each individual location.

- Local Indices for spatial Spatial Analysis (LISA)

Moran's I

This statistic is given by:

$$I = \frac{\sum_{i=1}^{n} \sum_{j=1, j \neq i}^{n} w_{ij} \left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{S_0 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2 / n} = \frac{n \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} \left(x_i - \bar{x}\right)\left(x_j - \bar{x}\right)}{S_0 \sum_{i=1}^{n} \left(x_i - \bar{x}\right)^2} \tag{15}$$

where $S_0 = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}$ and w_{ij} is an element of the spatial weight matrix that measures spatial distance or connectivity between regions i and j. In matrix form:

$$I = \frac{n}{S_0} \frac{\boldsymbol{z}^\top \boldsymbol{W} \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} \tag{16}$$

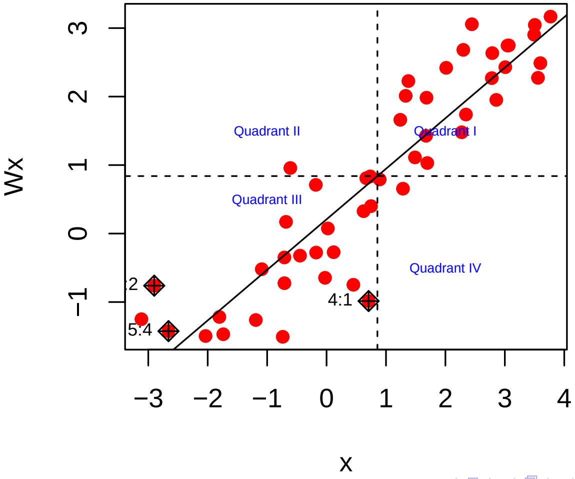
where $\boldsymbol{z} = \boldsymbol{x} - \bar{x}$. If the $\boldsymbol{W}$ matrix is row standardized, then:

$$I = \frac{\boldsymbol{z}^\top \boldsymbol{W}^s \boldsymbol{z}}{\boldsymbol{z}^\top \boldsymbol{z}} \tag{17}$$

because $S_0 = n$. Values range from -1 (perfect dispersion) to +1 (perfect correlation). A zero value indicates a random spatial pattern.

Moran Scatterplot

- A very useful tool for understanding the Moran's I test



Moran's I

Note that:

$$\widehat{\beta}_{OLS} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Therefore?

Moran's I

Note that:

$$\widehat{\beta}_{OLS} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

Therefore?

Remark

I is equivalent to the slope coefficient of a linear regression of the spatial lag $\boldsymbol{W}\boldsymbol{x}$ on the observation vector $\boldsymbol{x}$ measured in deviation from their means. It is, however, not equivalent to the slope of $\boldsymbol{x}$ on $\boldsymbol{W}\boldsymbol{x}$ which would be a more natural way.

Moran's I

- H_0: x is spatially independent, the observed x are assigned at random among locations. (I is close to zero)
- H_1: X is not spatially independent. (I is not zero)

Moran's I

- We are interested in the distribution of the following statistic:

$$T_I = \frac{I - \mathbb{E}(I)}{\sqrt{\mathbb{V}(I)}} \tag{18}$$

- Three approaches to compute the variance of Moran's I:
  - Monte Carlo
  - Normality of x_i: It is assumed that the random variable x_i are the result of n independently drawings from a normal population.
  - Randomization of x_i: No matter what the underlying distribution of the population, we consider the observed values of x_i were repeatedly randomly permuted.

Moran's I

Theorem (Moran's I Under Normality)

Assume that $\{\boldsymbol{x}_i\} = \{x_1, x_2, ..., x_n\}$ *are independent and distributed as* $\mathrm{N}(\mu, \sigma^2)$, *but* μ *and* σ^2 *are unknown. Then:*

$$\mathbb{E}\left(I\right) = -\frac{1}{n-1} \tag{19}$$

and

$$\mathbb{E}\left(I^2\right) = \frac{n^2 S_1 - n S_2 + 3 S_0^2}{S_0^2 (n^2 - 1)} \tag{20}$$

where $S_0 = \sum_{i=1}^n \sum_{j=1}^n w_{ij}$, $S_1 = \sum_{i=1}^n \sum_{j=1}^n (w_{ij} + w_{ji})^2 / 2$, $S_2 = \sum_{i=1}^n (w_{i.} + w_{.i})^2$, *where* $w_{i.} = \sum_{j=1}^n w_{ij}$ *and* $w_{i.} = \sum_{j=1}^n w_{ji}$ *Then:*

$$\mathbb{V}\left(I\right) = \mathbb{E}\left(I^2\right) - \mathbb{E}\left(I\right)^2 \tag{21}$$

Moran's I

Theorem 17 gives the moments of Moran's I under randomization.

Theorem (Moran's I Under Randomization)

Under permutation, we have:

$$\mathbb{E}\left(I\right) = -\frac{1}{n-1} \tag{22}$$

and

$$\mathbb{E}\left(I^2\right) = \frac{n\left[\left(n^2-3n+3\right)S_1 - nS_2 + 3S_0^2\right] - b_2\left[\left(n^2-n\right)S_1 - 2nS_2 + 6S_0^2\right]}{(n-1)(n-2)(n-3)S_0^2} \tag{23}$$

where $S_0 = \sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}$, $S_1 = \sum_{i=1}^{n}\sum_{j=1}^{n}(w_{ij}+w_{ji})^2/2$, $S_2 = \sum_{i=1}^{n}(w_{i.}+w_{.i})^2$, *where* $w_{i.} = \sum_{j=1}^{n} w_{ij}$ *and* $w_{i.} = \sum_{j=1}^{n} w_{ji}$*. Then:*

$$\mathbb{V}\left(I\right) = \mathbb{E}\left(I^2\right) - \mathbb{E}\left(I\right)^2 \tag{24}$$

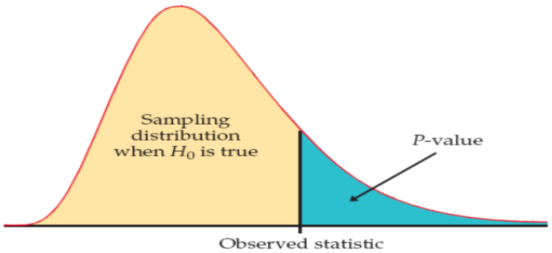
It is important to note that the expected value of Moran's I under normality and randomization is the same.

Monte Carlo

- Normality and randomization? We can use a Monte Carlo simulation
  - To test a null hypothesis H_0 we specify a test statistic T such that large values of T are evidence against H_0.
    - H_0 : no spatial autocorrelation.
  - Let T have observed value t_{obs}. We generally want to calculate:

$$p = \Pr(T \geq t_{obs} | H_0) \tag{25}$$

  - We need the distribution of T when H_0 is true to evaluate this probability.


Sampling
distribution
when H_0 is true
P-value
Observed statistic

Monte Carlo

Theorem (Moran's' I Monte Carlo Test)

The procedure is the following:

1. *Rearrange the spatial data by shuffling their location and compute the Moran's I S times. This will create the distribution under H_0.*
2. *Let $I_1^*, I_2^*, ..., I_S^*$ be the Moran's I for each time. A consistent Monte Carlo p-value is then:*

$$\widehat{p} = \frac{1 + \sum_{s=1}^{S} 1(I_s^* \geq I_{obs})}{S+1} \tag{26}$$

3. *For tests at the α level or at $100(1-\alpha)\%$ confidence intervals, there are reasons for choosing S so that $\alpha(S+1)$ is an integer. For example, use $S = 999$ for confidence intervals and hypothesis tests when $\alpha = 0.05$.*

Inference

Inference:

- If $I > \mathbb{E}(I)$, then a spatial unit tends to be connected by locations with similar attributes: Spatial clustering (low/low or high/high). The strength of positive spatial autocorrelation tends to increase with $I - \mathbb{E}(I)$.
- If $I < \mathbb{E}(I)$ observations will tend to have dissimilar values from their neighbors: Negative spatial autocorrelation (low/high or high/low)

Application

- `Lab1A.R`
- `Lab1B.R`