MBA USP ESALQ

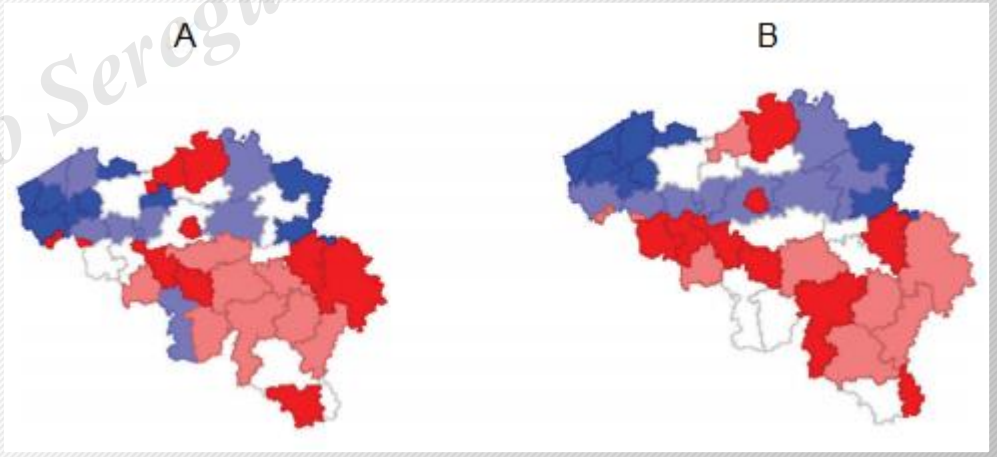
Spatial Analysis III

Rafael de Freitas Souza

Introduction to Exploratory Analysis of Spatial Data

Exploratory Analysis of Spatial Data

- According to Piatkowska, Messner and Yang (2018) the *Exploratory Spatial Data Analysis* (ESDA) is aimed at facilitating the identification and visualization of spatial patterns of a given phenomenon.



The evolution of criminality in Belgium between 2006 to 2012.
Source: Piatkowska et al. (2018), with adaptations.

How was it possible to evidence the considerations regarding the previous map?

The construction of the discussed maps was possible, firstly, because a **criterion of neighborhood** was defined. After this, a spatial **lag matrix W** was proposed and, only then, **spatial autocorrelations of the phenomenon were verified, in a global and local way**. After this, spatial models can be estimated.

Therefore:

- Step 1: Choose a criterion of neighborhood;
- Step 2: Build a matrix of spatial lag (**W**);
- Step 3: Calculate global and local autocorrelations;
- Step 4: Estimate models.

Establishing Neighborhoods

Spatial Lag Matrices

Neighborhood Matrices

The first step to perform a ESDA is to establish neighborhoods between the studied localities so that we can verify spatial autocorrelations, point out some heterogeneities, and even detect possible eventual outliers.

According to Anselin and Rey (2014) the establishment of neighborhoods is performed by a spatial weighing matrix **W** that can assume several types, being the most common the contiguity matrix, the geographical proximity and the socioeconomic proximity.

Matrix **W** Spatially Weighing by Contiguity

The idea of contiguity is from the assumption of the existence of a common physical border between spatial units. In this line of reasoning, it can be stated that Brazil is contiguous to Argentina, but it's not to China, for example.

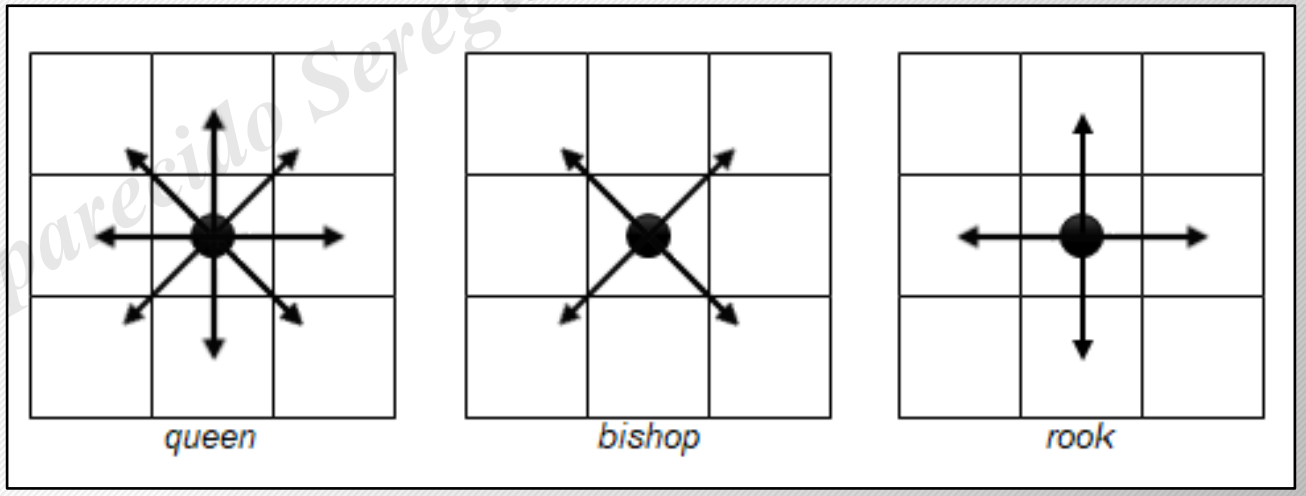
Based on the previously assumption, a Contiguity Matrix is a matrix **W**with binary values in which the value 1 is determined in the presence of a common physical border, and 0 for absence. We can describe your terms mathematically w_{ij}as:

$$w_{ij} = \begin{cases} 1, if\ there\ is\ contiguity\ between\ i\ e\ j; \\ 0\ , if\ there\ is\ not\ contiguity\ between\ i\ e\ j. \end{cases}$$

in which $w_{ii} = 0$,by convention.

Contiguity

The contiguity between elements w_{ij} can be established by several conventions. The most common conventions are the *queen*, the *bishop*, and the *rook*. The names are not coincidences and refer to the stipulated contiguities based on the movement of the chess pieces.



Fávero et al. (2022)

WSpatial Weighing Matrix by Geographic Distances

- The criterion of geographical distance can be an interesting option to promote the balance of neighbor between observations. From this perspective, it must be established a distance $d_i(k)$that is the threshold for establishing the $k$neighborhood to a given $i$region. According to Almeida (2012), the idea of this type of matrix is the premise that geographically closer regions have a greater spatial interaction.
- For the geographic distance criterion, the elements of $w_{ij}$the matrix $\mathbf{W}$can be mathematically presented by:

$$w_i(k) = \begin{cases} 1, for\ the\ cases\ in\ which\ d_{ij} \leq d_i(k); \\ 0, for\ the\ cases\ in\ which\ d_{ij} > d_i(k). \end{cases}$$

in which $w_{ii} = 0$,by convention.

WSpatial Weighing Matrix by$k$ -*Nearest Neighbors*

Anselin and Rey (2014) discuss that the weight W matrix by the k*nearest neighbors* convention doesn't make the researcher responsible for deciding an optimal distance $d_i(k)$between the observations of their database, in addition to avoid the existence of “islands" on the map. In opposition, in our opinion, it requires the researcher to decide on the value of $k$neighbors. it is also necessary to say that it is, once $w_{ii} = 0$ again, by convention.

WSpatial Weighing Matrix by Social Distances

We can also establish a matrix W using social distances as a criterion of spatial weight criterion. We can understand social distances as the differences of IDH, of Gini's Index , of illiteracy rates, of child mortality rates, etc. between the observations.

Spatial Matrices Standardization

Spatial Matrices Standardization

Anselin and Rey (2014) state that the use of a matrix **W** in its binary form is rare, suggesting to adopt some process of standardization, which the most common are **W** row-standardization, double standardization and variance stabilizing.

Suppose the following matrix **W**

	ARG	BOL	BRA	CHL	COL	ECU	GUY	GUF	PRY	PER	SUR	URY	VEN
ARG	0	1	1	1	0	0	0	0	1	0	0	1	0
BOL	1	0	1	1	0	0	0	0	1	1	0	0	0
BRA	1	1	0	0	1	0	1	1	1	1	0	0	0
CHL	1	1	0	0	0	0	0	0	0	1	0	0	0
COL	0	0	1	0	0	1	0	0	0	1	0	0	1
ECU	0	0	0	0	1	0	0	0	0	1	0	0	0
GUY	0	0	1	0	0	0	0	0	0	0	1	0	1
GUF	0	0	1	0	0	0	0	0	0	0	1	0	0
PRY	1	1	1	0	0	0	0	0	0	0	0	0	0
PER	0	1	1	1	1	1	0	0	0	0	0	0	0
SUR	0	0	1	0	0	0	1	1	0	0	0	0	0
URY	1	0	1	0	0	0	0	0	0	0	0	0	0
VEN	0	0	1	0	1	0	1	0	0	0	0	0	0

Row-Standardization $\mathbf{W}$

Assuming the exposed by our table as example of neighborhood matrix, the row standardized matrix $\mathbf{W}$ considers the sum of the binary spatial weights in each of its rows, dividing them by their respectivew_{ij} . Mathematically:

$$w_{ij}{}_{\text{ row standardized}} = \frac{w_{ij}}{\sum_j w_{ij}}$$

in which the sum of the spatial weight of each row must be equal to 1; while the sum of all weight $S_0$is given by:

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

in which n it is equal to the total of observations; if there are q observations without neighbors, these must be subtracted from$,n,$ generating the exposed by:

$$S_0 = \sum_i \sum_j w_{ij} = n - q$$

Row-StandardizationW

	ARG	BOL	BRA	CHL	COL	ECU	GUY	GUF	PRY	PER	SUR	URY	VEN
ARG	0.00	0.20	0.20	0.20	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00
BOL	0.20	0.00	0.20	0.20	0.00	0.00	0.00	0.00	0.20	0.20	0.00	0.00	0.00
BRA	0.14	0.14	0.00	0.00	0.14	0.00	0.14	0.14	0.14	0.14	0.00	0.00	0.00
CHL	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
COL	0.00	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.25
ECU	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00
GUY	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.33
GUF	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00
PRY	0.33	0.33	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PER	0.00	0.20	0.20	0.20	0.20	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SUR	0.00	0.00	0.33	0.00	0.00	0.00	0.33	0.33	0.00	0.00	0.00	0.00	0.00
URY	0.50	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VEN	0.00	0.00	0.33	0.00	0.33	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00

Double Standardization W

The idea of the double standardization procedure of the spatial weight matrix W is to transform it into stochastic matrix, which sum of all your weight S_0 is equal to 1:

$$w_{ij}{}_{\text{Double standardized}} = \frac{w_{ij}}{\sum_i \sum_j w_{ij}}$$

	ARG	BOL	BRA	CHL	COL	ECU	GUY	GUF	PRY	PER	SUR	URY	VEN
ARG	0.00	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.02	0.00
BOL	0.02	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00
BRA	0.02	0.02	0.00	0.00	0.02	0.00	0.02	0.02	0.02	0.02	0.00	0.00	0.00
CHL	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
COL	0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.00	0.00	0.02
ECU	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.02	0.00	0.00	0.00
GUY	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.02
GUF	0.00	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	0.00	0.00
PRY	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PER	0.00	0.02	0.02	0.02	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SUR	0.00	0.00	0.02	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00
URY	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VEN	0.00	0.00	0.02	0.00	0.02	0.00	0.02	0.00	0.00	0.00	0.00	0.00	0.00

Matrix Standardization **W** by Stabilizing the Variance

The standardization of the spatial weight matrix W by stabilizing the variance was proposed by Tiefelsdorf, Griffith and Boots (1999). The standardized value of spatial weight is obtained in two steps.

First, each original weight of the row i must be divided by the square root of the sum of the weights squared of their respective i row, originating a new weight called w_{ij}^*:

$$w_{ij}^* = \frac{w_{ij}}{\sqrt{\sum_j w_{ij}^2}}$$

Hereupon, each weight w_{ij}^* must be multiplied by the factor presents in:

$$\frac{n-q}{Q}$$

in which n it is the total of observations; it q is the total of observations without neighbors; Q is given by $\sum_i \sum_j w_{ij}^*$

Matrix **W** Standardization by Stabilizing the Variance

	ARG	BOL	BRA	CHL	COL	ECU	GUY	GUF	PRY	PER	SUR	URY	VEN
ARG	0.00	0.24	0.24	0.24	0.00	0.00	0.00	0.00	0.24	0.00	0.00	0.24	0.00
BOL	0.24	0.00	0.24	0.24	0.00	0.00	0.00	0.00	0.24	0.24	0.00	0.00	0.00
BRA	0.20	0.20	0.00	0.00	0.20	0.00	0.20	0.20	0.20	0.20	0.00	0.00	0.00
CHL	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.00	0.00
COL	0.00	0.00	0.27	0.00	0.00	0.27	0.00	0.00	0.00	0.27	0.00	0.00	0.27
ECU	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00
GUY	0.00	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.31	0.00	0.31
GUF	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00
PRY	0.31	0.31	0.31	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
PER	0.00	0.24	0.24	0.24	0.24	0.24	0.00	0.00	0.00	0.00	0.00	0.00	0.00
SUR	0.00	0.00	0.31	0.00	0.00	0.00	0.31	0.31	0.00	0.00	0.00	0.00	0.00
URY	0.38	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
VEN	0.00	0.00	0.31	0.00	0.31	0.00	0.31	0.00	0.00	0.00	0.00	0.00	0.00

Spatial Autocorrelation

Spatial Autocorrelation

After establishing the neighborhoods, and their respective **W** matrix of spatial lag, we can verify if the observed data are distributed in a random way or it is a spatial pattern, that is, **if there is the spatial autocorrelation involving the studied phenomenon**.

Griffith (2003) states that spatial autocorrelation can be understood as the measure of the existing correlation between the values of a single variable of interest in a geographical way.

- The global autocorrelation metrics are directed to measure the degree of spatial relation of a phenomenon regarding all values observed in the database.
- In opposition, the local autocorrelation metrics measure **the** autocorrelations of observations, one by one, regarding their neighborhood established by the spatial lag matrix.

I Global Autocorrelation – the Moran Statistics

The statistics I was first proposed by Moran (1948) and, years later, Cliff and Ord (1973, 1981) presented a more developed work on the original ideas of Moran, determining the following formula:

$$I = \frac{n}{S_0} \times \frac{\sum_i \sum_j w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

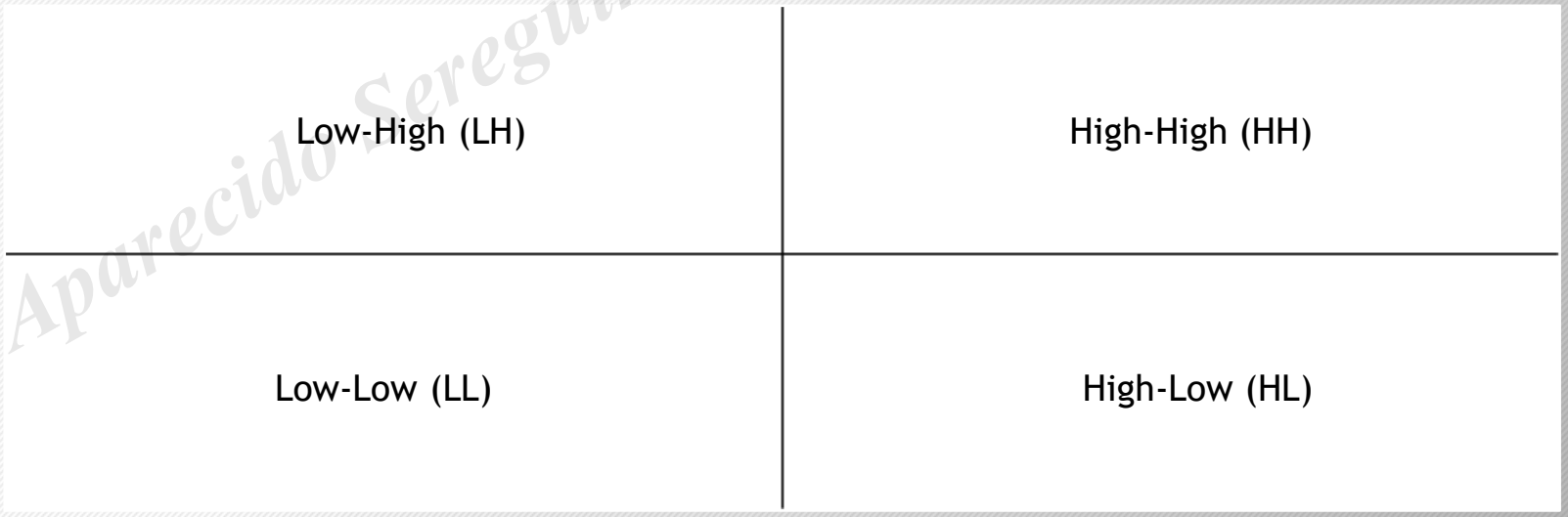
in which n it represents the number of observations; z points out the standardized values of the dependent variable Y by the procedure $zscores$; w_{ij} represent the spatial weight of the matrix **W** of a given observation in the row i and in the column j; and S_0 it is the sum of all spatial weights w_{ij}.

$H_0: -\left[\frac{1}{n-1}\right]$ which indicates that the values Y_i are independent of the values of neighboring observations, that is, that there is no spatial autocorrelation for a given significance level.

Moran's Diagram of Statistics I

The Moran's Diagram is a method of visualize global autocorrelations indicated by the Moran's *I* statistics.

The visual technique consists of a 2D graph of dispersion with four quadrants, namely: High-High (HH), Low-Low (LL), High-Low (HL) and Low-High (LH):



Local Autocorrelation – the Local Moran Statistic

Anselin (1995) proposed a way to measure the local autocorrelations called *Local Indicators of Spatial Association* (LISA). The LISA technique aims to identify local patterns of spatial association (Anselin, 1995, p. 93).

According to Lansley and Cheshire (2016), the LISA technique investigates the spatial relations between the data, considering the established neighborhoods. Among the types of LISA proposed by Anselin (1995) – e.g. Gamma Local, Geary Local, etc. – the most commonly referenced type will be presented: the Local Moran.

$$I_i = z_i \sum_j w_{ij} z_j$$

in which, in a similar way to Moran's I Statistics, z_i and z_j represent the standardized values of the dependent variable; the considered sum includes only each neighbor that j belongs to the neighborhood J_i established by the spatial weighing W matrix. and the spatial weight w_{ij} are, preferentially, row standardized to facilitate the interpretation, without omitting that, by convention, $w_{ii} = 0$.

Local Autocorrelation - Getis-Ord General G

Getis and Ord (1992) proposed another way to study the spatial association of observations of a given database, based on the spatial concentration.

G_iAccording to Almeida (2012) the statistics can select a metric for each observation that determines in which measure the individuals of the database are surrounded by observations with high values – called as *hot spots*; or surrounded by observations with low values – called as *cool spots*.

$G_i(d) = \frac{\sum_{j-1}^{n} w_{ij(d)} Y_j}{\sum_{j-1}^{n} Y_j}$, where $j \neq i$

in which w_{ij} it represents the binary weight of a spatial weighing matrix by distances and, by convention, $w_{ii} = 0$;the numerator represents the sum of all neighboring values Y_j within the neighborhood established by the distance d of i, without Y_i ; and the denominator represents the sum of all neighboring values Y_j without Y_i.

References

Almeida, E. (2012).
Econometria Espacial Aplicada. Campinas: Alínea.

Anselin, L. (1995). Local Indicators of Spatial Association. *Geographical Analysis, 27*(2), 93-115. doi:10.1111/j.1538-4632.1995.tb00338.x

Anselin, L., & Rey, S. J. (2014). *Modern Spatial Econometrics in Practice*. Chicago: GeoDa Press.

Cliff, A. D., & Ord, J. K. (1973). Classics in Human Geography Revisited. *Progress in Human Geography, 19*(2), 245-249. doi:10.1177/030913259501902020 5

Cliff, A. D., & Ord, J. K. (1981). *Spatial Processes Models and Applications*. London: Pion.

Fávero, L. P., Belfiore, P., & Freitas Souza, R. (2022). Data Science, Analytics and Machine Learning with R. Cambridge: Academic Press.

Getis, A., & Ord, J. K. (1992). The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis, 24*(3), 189-206. 1Geographical Analysis, 241(3), 189-206. doi:10.1111/j.1538-4632.1992.tb00261.x

Griffith, D. A. (2003). *Spatial Autocorrelation and Spatial Filtering: Gaining Understanding Through Theory and Scientific Visualization (Advances in Spatial Science)*. London: Springer.

Piatkowska, S. J., Messner, S. F., & Yang, T. -C. (2018). Xenophobic and racially motivated crime in Belgium: exploratory spatial analysis and spatial regressions of structural covariates. *Deviant Behavior, 39*(11), 1398-1418. doi:10.1080/01639625.2018.1479917

Tiefelsdorf, M., Griffith, D. A., & Boots, B. (1999). A Variance-Stabilizing Coding Scheme for Spatial Link Matrices. *Environment and Planning A, 31*(1), 165-180. doi:10.1068/a310165