Phd Program in Transportation

Transport Demand Modeling

Anabela Ribeiro - 2021
anabela@dec.uc.pt

Spatial Regression Models

General Framework

Class Structure

	Dia	Description
1	17 November 2021 (3 hours)	A - Panel data models: Main issues One way component error models Two way component error models *Exercises and notes on the Home Assignment* *Panel Data Models* B – Spatial regression models: Main issues Exploratory Spatial Data Analysis
2	19 November 2021 (3 hours)	B – Spatial regression models: Spatial Regression Models *Exercises and notes on the Home Assignment* *Spatial Data Models*

Spatial Data Models

Main issues

Exploratory Spatial Data Analysis

Spatial Regression analysis

Spatial Data Models: General Presentation

Near things tend to be more related than faraway things (Aldo Tobler)

Spatial autocorrelation considers the influence of proximity in the definition of clusters and outliers.

Spatial Analysis: Statistical analysis for the description and modeling of processes going on in space and where the 'spatial factor' may have an important explanatory role. There are two main types:

- **Exploratory spatial data analysis** (spatial autocorrelation with statistical measures)
- **Modeling spatial data analysis** (regression analysis)

INSTITUTO SUPERIOR TÉCNICO

FEUP

Spatial Data Models: General Presentation

- **Concept of spatial autocorrelation** – what happen in an unit relates with what happens in its neighbors

↓

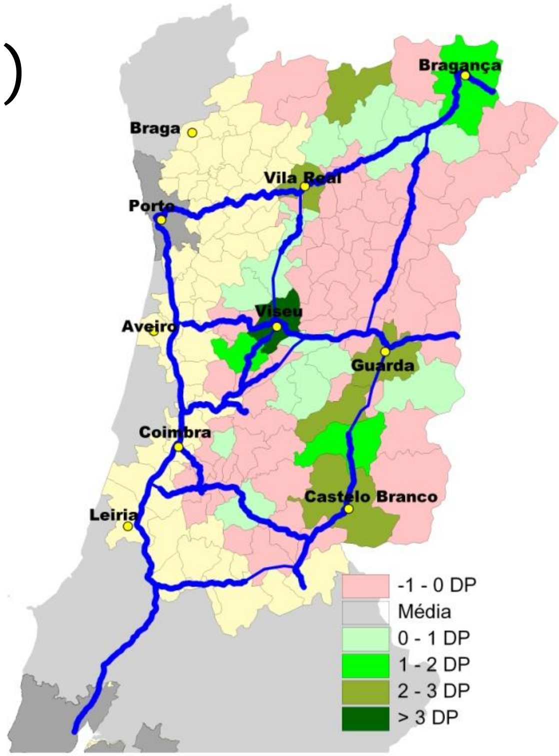
- **Concept of spatial neighbor** – defined exogenously

↓

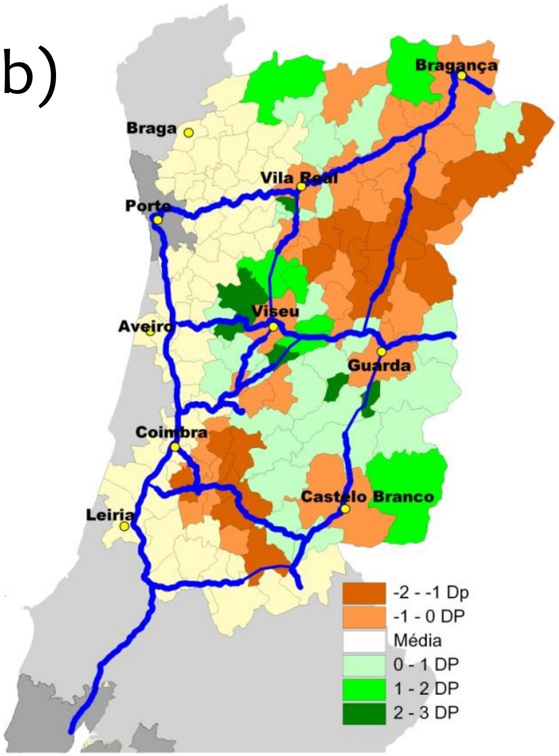
- **Concept of spatial structure** – defined through the definition of spatial neighboring

Spatial Data Models: General Presentation

(a)



(b)



Population distribution without spatial effects (a)

Population distribution with spatial effects – spatial lag (b)

(Weighted average of the values in the unit and in the neighbor units)

Spatial data models applications

Type of applications:

- Economic convergence and agglomeration economies
- Urban development
- New transport infrastructures impact on development
- Hedonic prices
- Four step transport model

Spatial data models applications

Study the impact of new road infrastructures when the accessibility of city centers is one of the independent variables

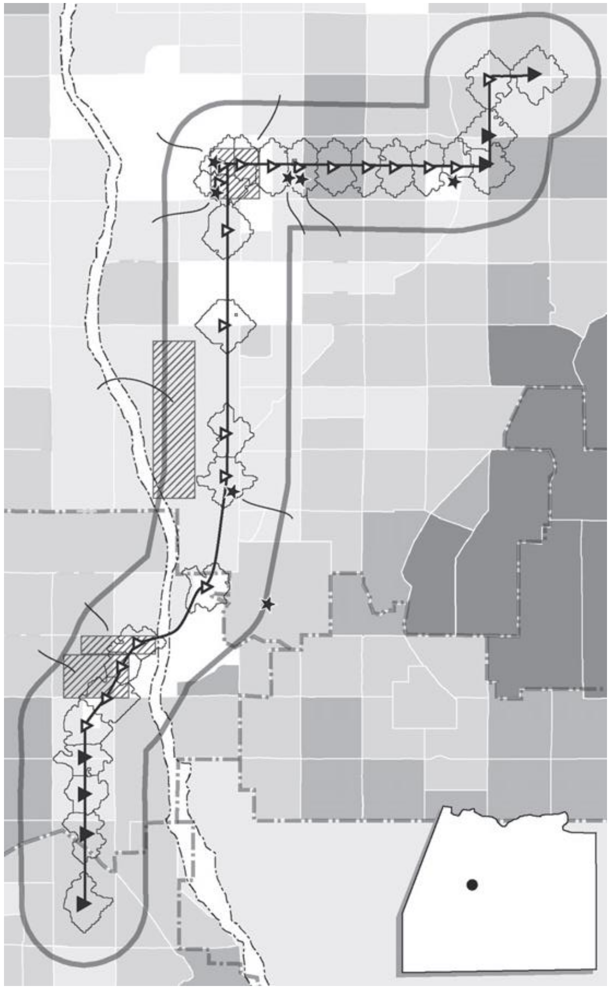
Does the position relatively to the network is influencing local development?

How?

Spatial data models applications

Study of the impacts of
new infrastructure of public
transportation in land prices
TOD Strategies

Is there a relationship between
the accessibility to a station and
the value of houses?
Is there a significant relationship
between the geographical units?



Spatial data

Data types

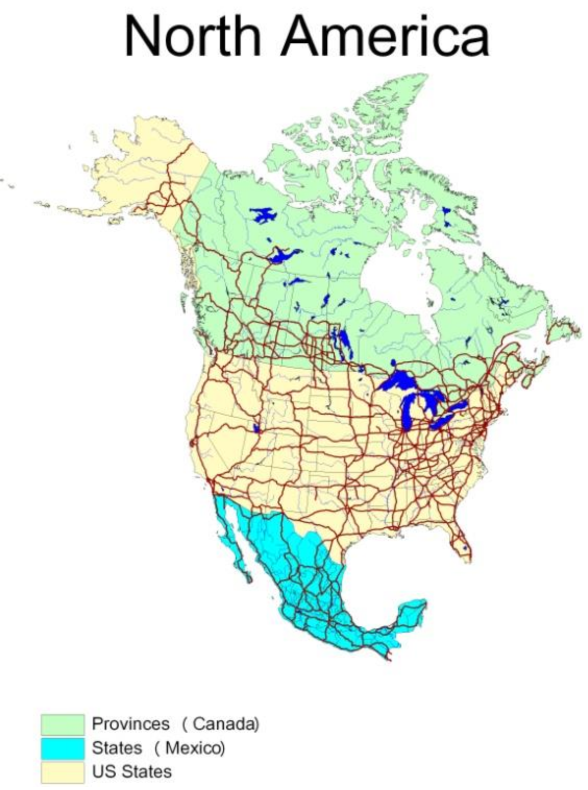
Point patterns (ex. Epidemic – focus of diseases)

Spatially continuous data (ex. Geostatistics – land use)

Spatial areal data (ex. Economics – census data)

Interaction (flows)

Spatial data - areas



AREA	CODE	NAME	POP1990	POP90_SQMI	P_URBAN90	P_ING_LANG	P_EMPL_SEC	HSE_UNIT90
28002,325	MX02	Baja California Norte	1660855	61,537060	90,904620	1,300000	31,700000	362727
27898,191	MX03	Baja California Sur	317764	11,204120	78,254620	1,000000	18,800000	67304
10547,762	MX18	Nayarit	824643	79,186840	62,054850	3,400000	17,600000	168451
30736,386	MX14	Jalisco	5302689	169,943300	81,853420	0,500000	32,700000	1029178
2110,761	MX01	Aguascalientes	719659	340,778900	76,521930	0,100000	34,200000	129853
11715,793	MX11	Guanajuato	3982593	338,381600	63,414290	0,300000	35,000000	687136
4645,565	MX22	Queretaro de Arteaga	1051235	237,872900	59,723950	2,300000	37,300000	193434
8198,684	MX13	Hidalgo	1888366	235,052100	44,785700	19,500000	25,200000	362933
22961,931	MX16	Michoacan de Ocampo	3548199	153,387800	61,618700	3,500000	23,200000	663496
8291,708	MX15	Mexico	9815795	1190,799000	84,406890	3,700000	36,800000	1876545
512,050	MX09	Distrito Federal	8235744	14426,050000	99,734080	1,500000	27,000000	1789171
2211,033	MX08	Colima	428510	213,856600	83,319880	0,400000	21,400000	88627
1951,676	MX17	Morelos	1195059	625,456100	85,621540	1,900000	27,900000	244958
14628,106	MX31	Yucatan	1362940	91,946590	78,625470	44,200000	24,500000	273958
19369,415	MX04	Campeche	535185	27,286660	70,024380	19,000000	19,400000	107894
13227,318	MX21	Puebla	4126101	315,302400	64,292630	14,100000	24,900000	772461
19782,748	MX23	Quintana Roo	493277	25,450490	73,868030	32,200000	15,500000	102859
1534,069	MX29	Tlaxcala	761277	491,090700	76,496600	3,400000	33,900000	137135
25001,188	MX12	Guerrero	2620637	105,617700	52,259660	13,400000	16,900000	501725
35786,813	MX20	Oaxaca	3019560	83,262670	39,452870	39,100000	16,400000	587131
9364,727	MX27	Tabasco	1501744	153,976700	49,656800	3,700000	20,500000	285319
28335,571	MX05	Chiapas	3210496	112,077000	40,390700	26,400000	11,100000	594025
69542,455	MX26	Sonora	1823606	25,950650	79,132610	3,000000	25,400000	378587
95771,458	MX06	Chihuahua	2441873	25,827340	77,390020	2,900000	35,900000	529799
57988,415	MX07	Coahuila De Zaragoza	1972340	34,068680	86,056210	0,200000	37,700000	404691
22315,678	MX25	Sinaloa	2204054	97,894390	64,084050	1,600000	17,200000	422242
46463,611	MX10	Durango	1349378	28,379360	57,390660	1,600000	26,600000	262164
28881,617	MX32	Zacatecas	1276323	45,139200	45,937980	0,100000	21,500000	238779
24767,049	MX24	San Luis Potosi	2003187	82,285840	55,163250	11,900000	26,000000	379336
25139,835	MX19	Nuevo Leon	3098736	123,649400	91,994190	0,200000	40,200000	642298
30503,273	MX28	Tamaulipas	2249581	73,414430	81,068610	0,400000	30,500000	488508
27564,808	MX30	Veracruz-Llave	6228239	225,042700	56,223370	10,700000	21,200000	1262509

Spatial data – point patterns

NAME	CAPITAL	STATE_NAME	POPULATION
Monterrey	Y	Nuevo Leon	2015000
Mazatlan	N	Sinaloa	199830
Guadalajara	Y	Jalisco	2325000
Tampico	N	Tamaulipas	435000
Mexico City	C	Distrito Federal	14100000
Puebla de Zaragoza	Y	Puebla	1055000
Veracruz	N	Veracruz-Llave	385000
Oaxaca	Y	Oaxaca	154223
Merida	Y	Yucatan	400142
Mexicali	Y	Baja California Norte	365000
Aguascalientes	Y	Aguascalientes	293152
Campeche	Y	Campeche	128434
La Paz	Y	Baja California Sur	91453
Tuxtla Gutierrez	Y	Chiapas	131096
Chihuahua	Y	Chihuahua	385603
Saltillo	Y	Coahuila De Zaragoza	284937
Colima	Y	Colima	86044
Durango	Y	Durango	257915
Guanajuato	Y	Guanajuato	48981
Chilpancingo	Y	Guerrero	67498
Pachuca	Y	Hidalgo	110351
Morelia	Y	Michoacan de Ocampo	297544
Toluca	Y	Mexico	199778
Cuernavaca	Y	Morelos	192770
Tepic	Y	Nayarit	145741
Queretaro	Y	Queretaro de Arteaga	215976
Chetumal	Y	Quintana Roo	56709
San Luis Potosi	Y	San Luis Potosi	470000
Culiacan	Y	Sinaloa	304826
Hermosillo	Y	Sonora	297175
Villahermosa	Y	Tabasco	158216
Tlaxcala	Y	Tlaxcala	35384
Jalapa	Y	Veracruz-Llave	204594
Zacatecas	Y	Zacatecas	80088
Ciudad Victoria	Y	Tamaulipas	140161
Acapulco	N	Guerrero	301902

Spatial Autocorrelation

Spatial data exploration (**spatial autocorrelation** with statistical measures) in which the space is included. It helps in the identification of some trends in space behavior in the variables that are part of a particular space model.

Spatial data modeling (**spatial regression analysis**) specification and estimation methods to investigate the significance of transportation accessibility (for example) as an explanatory factor in spatial models.

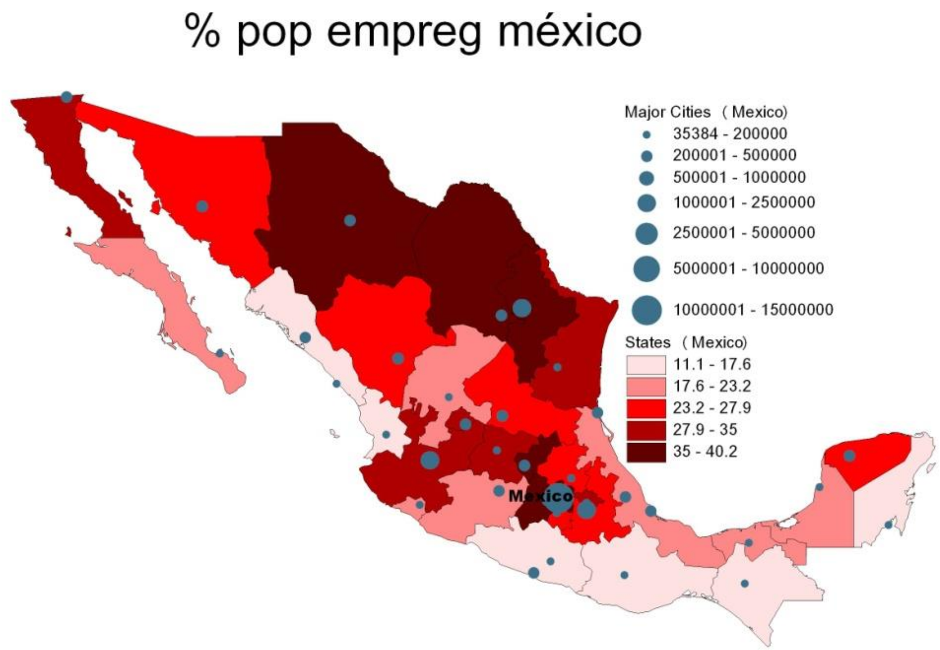
Spatial Autocorrelation

Is there any spatial correlation for unemployment in Mexico?

Can we define spatial tendencies and/or clusters or outliers, with **statistical significance?**

↓

Development strategies.



Spatial Autocorrelation

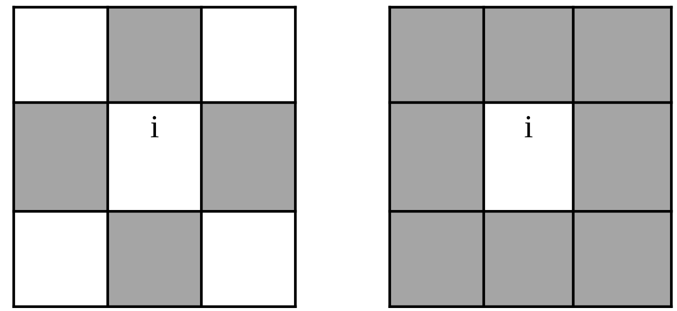
1st step: defining spatial matrices

- Defining a spatial structure that is a spatial relation of proximity between all the units in the area under study.

- The spatial structure of the neighborhood for a set of ***n*** is defined by territorial units of a square matrix ***W (nxn)*** which indicates the relationship between each pair of territorial units ***i*** and ***j***. Each cell of the matrix is called ***w_ij***, whose value is determined by the criteria used to define the relation of proximity between ***i*** and ***j.***

Spatial Autocorrelation

Spatial matrices – criteria for neighbors relations structure



Rook Queen

Two different ways of defining neighbors

…….

Simplest criteria:

w_ij = 1 if i shares a border line with j

w_ij = 0 if i does not share a border line with j

Spatial Autocorrelation

Spatial matrices – criteria for neighbors relations structure

Other criteria:

w_ij = 1 if j is within a distance of <d

w_ij = 0 if j is within a distance of >=d

w_ij = 1 if j is one of the K-nearest neighbors

w_ij = 0 if j is not one of the K-nearest neighbors

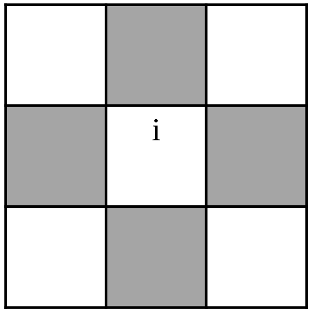
…..

Spatial Autocorrelation

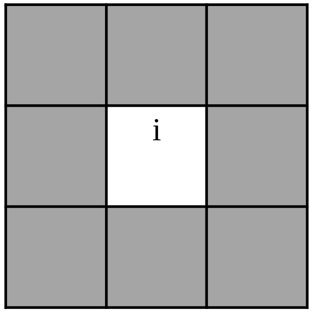
Spatial matrices

Z1	Z2	Z3
Z4	Z5	Z6
Z7	Z8	Z9

Spatial area with 9 units



Rook



Queen

W_{25}

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9
Z1	0	1	0	1	0	0	0	0	0
Z2	1	0	1	0	1	0	0	0	0
Z3	0	1	0	0	0	1	0	0	0
Z4	1	0	0	0	1	0	1	0	0
Z5	0	**1**	0	**1**	0	**1**	0	**1**	0
Z6	0	0	1	0	1	0	0	0	1
Z7	0	0	0	1	0	0	0	1	0
Z8	0	0	0	0	1	0	1	0	1
Z9	0	0	0	0	0	1	0	1	0

Matrix W:

- *rook*
- 1ºst neighbors

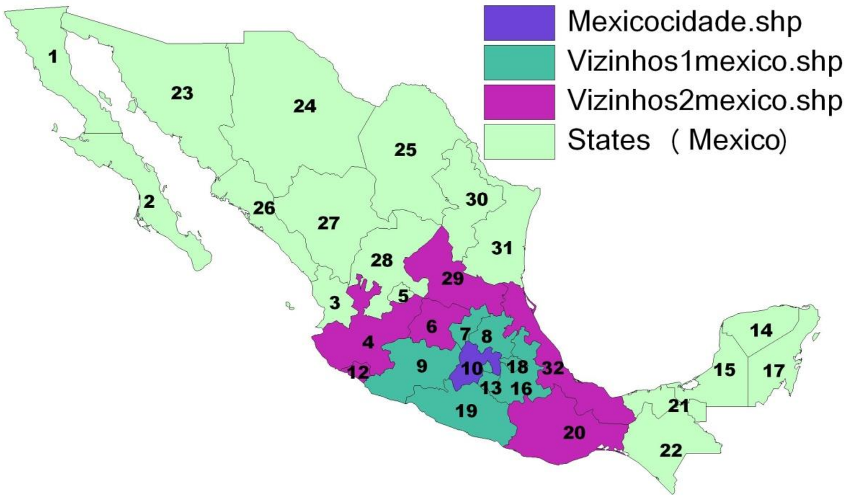
Spatial Autocorrelation

Matrix W (1) - Spatial neighbor structure in (‘1st neighbors);

Matrix W(2) - Spatial neighbor structure in (‘2nd neighbors’) this is to say neighbors that are closer after the first neighbors.

Etc…

Mexico territorial structure



Spatial Autocorrelation

Global and local autocorrelation

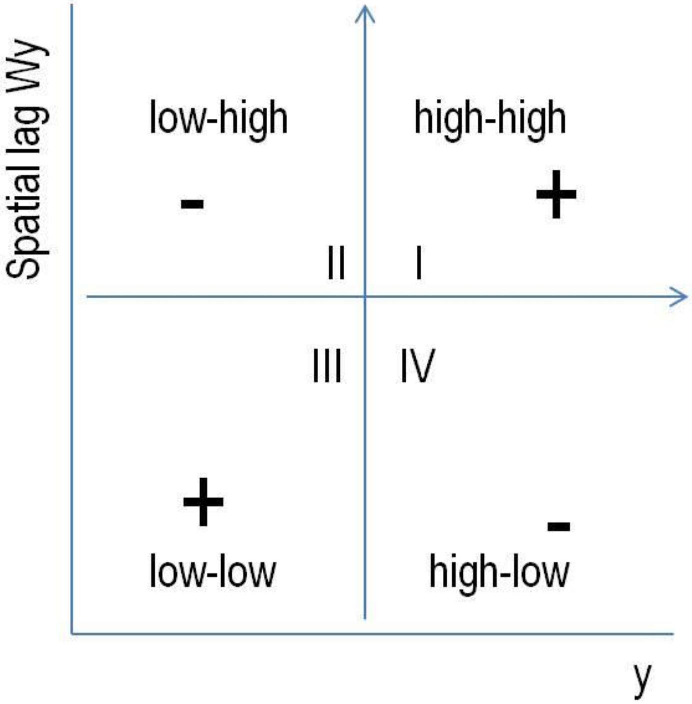
- **Global autocorrelation** – Autocorrelation for a variable considering all the space under analysis

- **Local autocorrelation** - Autocorrelation for a variable for a specific spatial unit but considering all the spatial units that are its neighbors

Spatial Autocorrelation

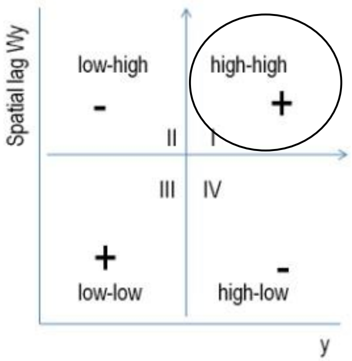
Wy - Spatial lag of a variable – average of the values in a certain spatial unit and in its **neighbors**.

$$Wy = \frac{\sum_{j=1}^{n} w_{ij} y_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

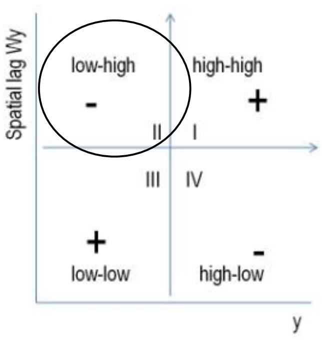


Spatial Autocorrelation

1st sector (I) – Positive relation (+) high values (for the variable) surrounded by high values (for the average in the neighbors) – high-high clusters;

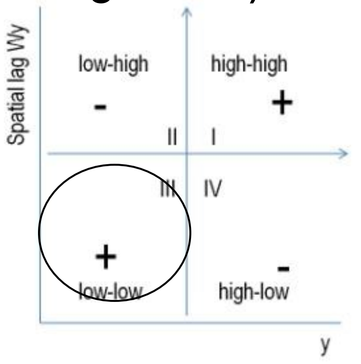


2nd sector (II) – Negative relation (-) low values (for the variable) surrounded by high values (for the average in the neighbors) – low -high outliers;

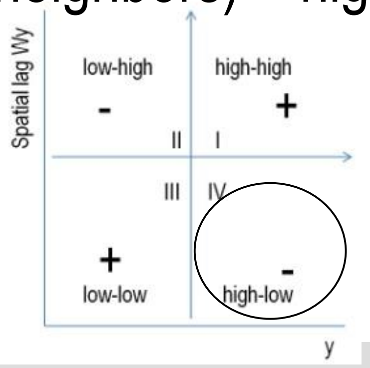


Spatial Autocorrelation

3rd sector (III) – Positive relation (+) low values (for the variable) surrounded by low values (for the average in the neighbors) – low-low clusters;



4rd sector (IV) – Negative relation (-) high values (for the variable) surrounded by low values (for the average in the neighbors) – high - low outliers;



Spatial Autocorrelation

Global Autocorrelation – *Moran's I*: slope for a linear regression adjusting all the points.

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

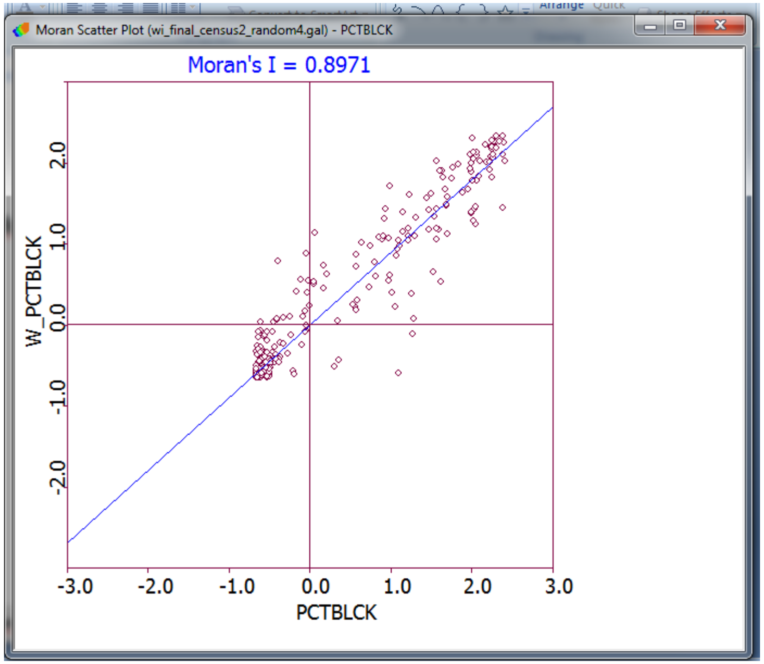
Local Autocorrelation – autocorrelation for each spatial unit.

$$I_i = \frac{(y_i - \bar{y})}{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{N}} \sum_i \sum_{i=1}^{n} w_{ij}(y_i - \bar{y}) \qquad I = \sum_i \frac{I_i}{N}$$

Spatial Autocorrelation

Global Autocorrelation – *Moran's I*: slope for a linear regression adjusting all the points. Indicator of a global autocorrelation trend for all the space under analysis.

(Values for *y* an *Wy* in deviations from the mean)

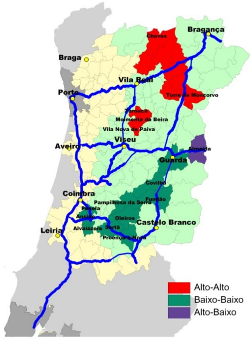


Spatial Autocorrelation

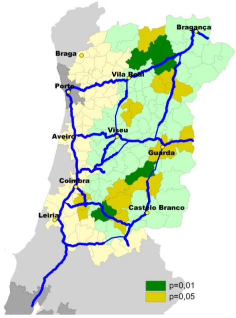
Local Autocorrelation – autocorrelation for each spatial unit.

LISA maps – Local I Spatial Autocorrelation

Cluster map

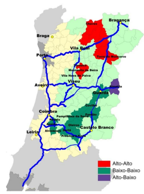


Significance map



Spatial Autocorrelation

Cluster map – indicating units with statistical significance for spatial autocorrelation



high-high clusters - high values (for the variable) surrounded by high values (for the average in the neighbors)

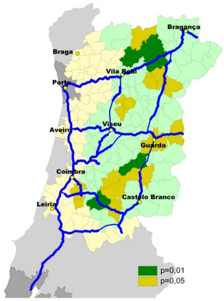
low-high outliers - low values (for the variable) surrounded by high values (for the average in the neighbors)

low-low clusters - low values (for the variable) surrounded by low values (for the average in the neighbors)

high-low outliers - high values (for the variable) surrounded by low values (for the average in the neighbors) – high - low outliers;

Spatial Autocorrelation

Significance map

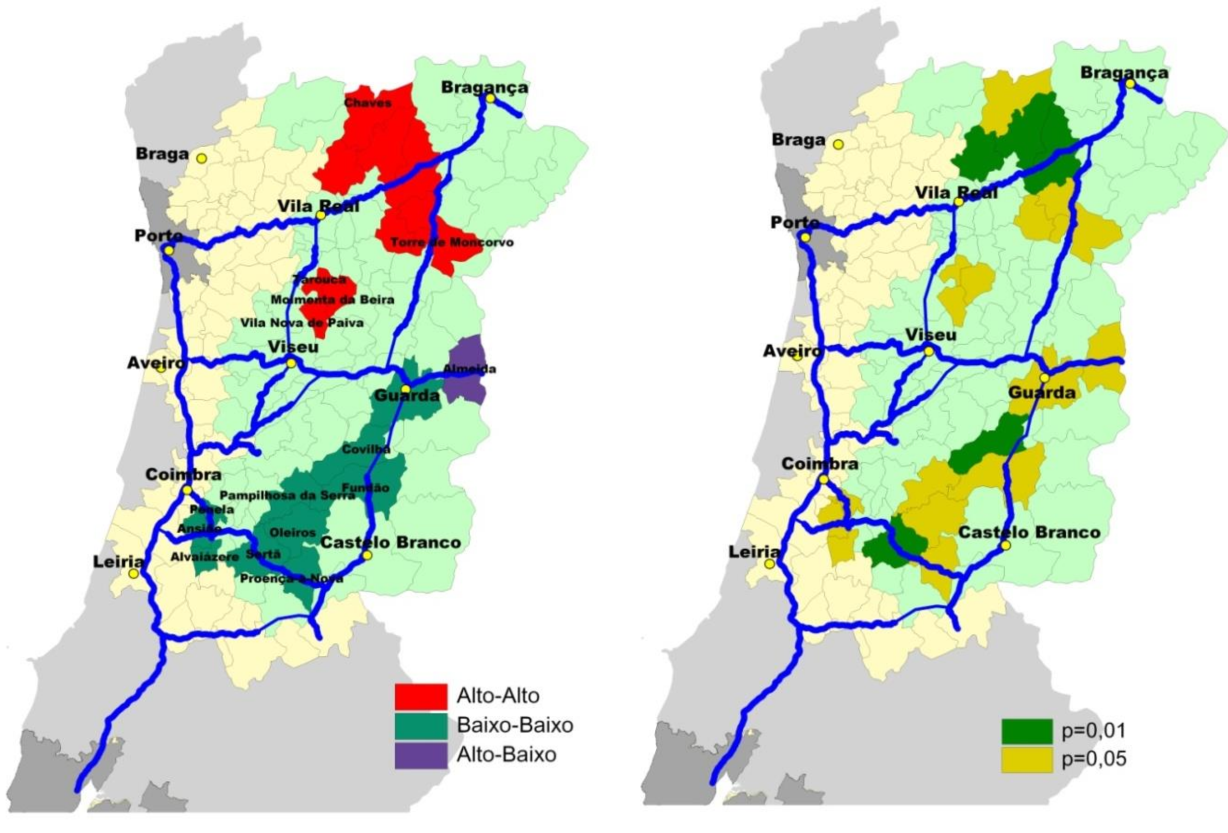


The map of statistical significance is synchronized with the previous one and shows the degree of significance from a p value of 0.05, i.e. the situation where one can reject with 95% probability the null hypothesis of no autocorrelation (p = 0.01 already implies that this hypothesis can be rejected with 99% probability)

H0: There is no autocorrelation between a unit and the neighboring
H1: There is autocorrelation

Spatial Autocorrelation

Clusters map(a) statistical significance map(b) for unemployment in 2001



Spatial Autocorrelation – Example 1

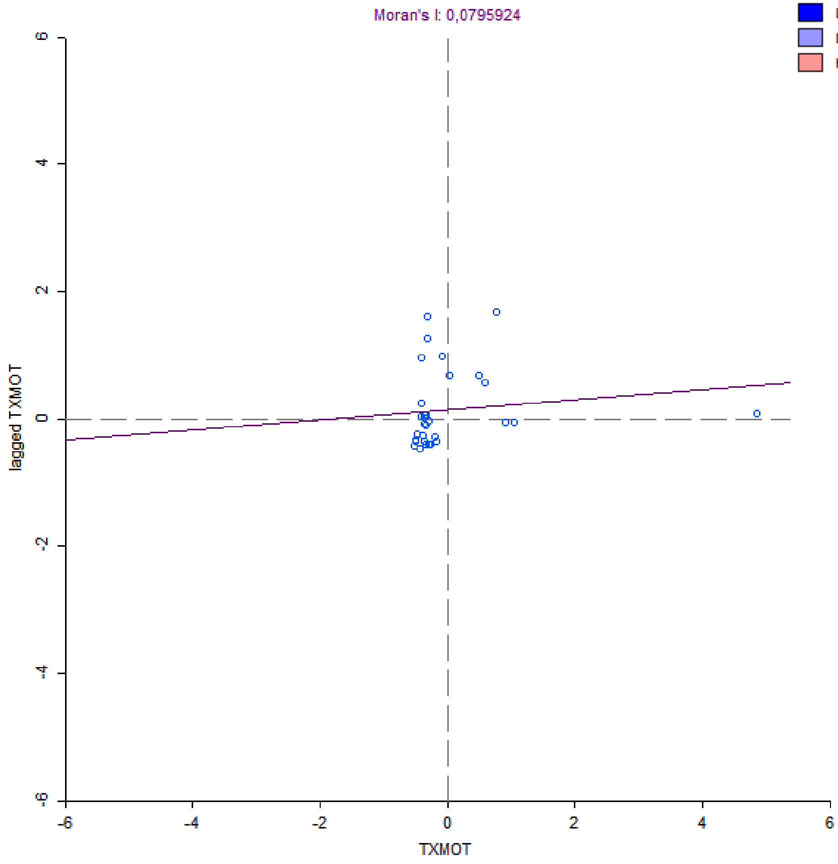
Download Geoda

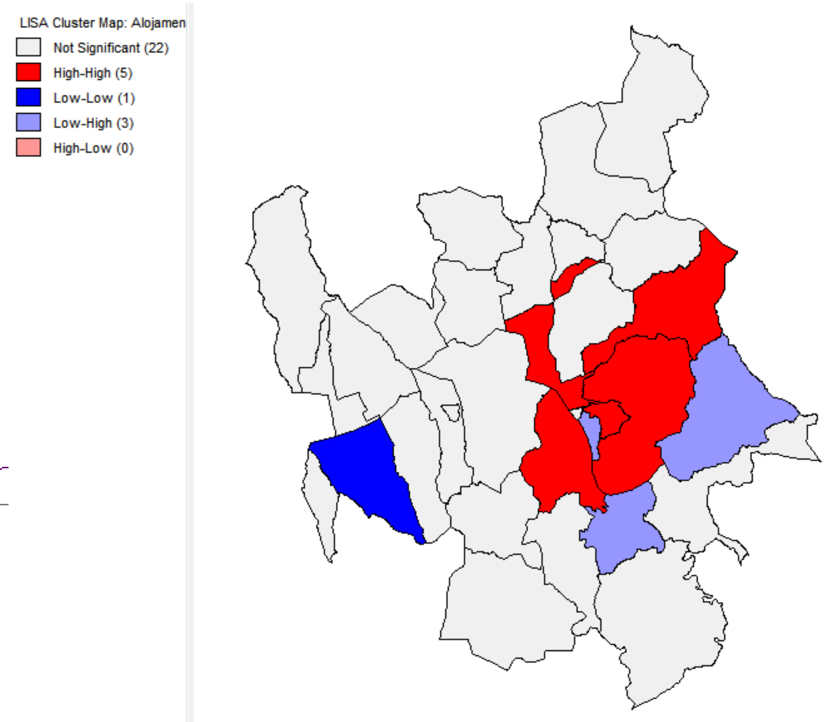
Input in Geoda the data base CensosCoimbra2001_alojamentos

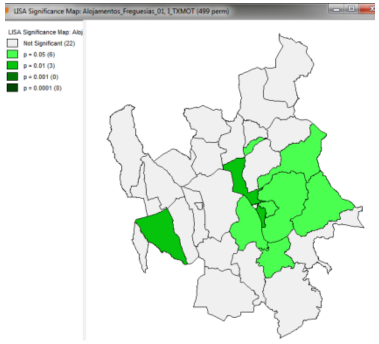
Create a new variable similar to motorization rate, assuming that to each ‘alojamento’ corresponds 1,5 cars.

Present and analyse the Moran I scatterplot , the Cluster Map and the Significance Map for this motorization rate.

Spatial Autocorrelation – Example 1







Spatial Autocorrelation – Example 2

Box Map

1. Open the Milwaukee shapefile - wi_final_census2_random4 (use *FIPS as your key variable)*
2. Create a box map (1.5 hinge) with variable *PCTBLCK*

LISA Map

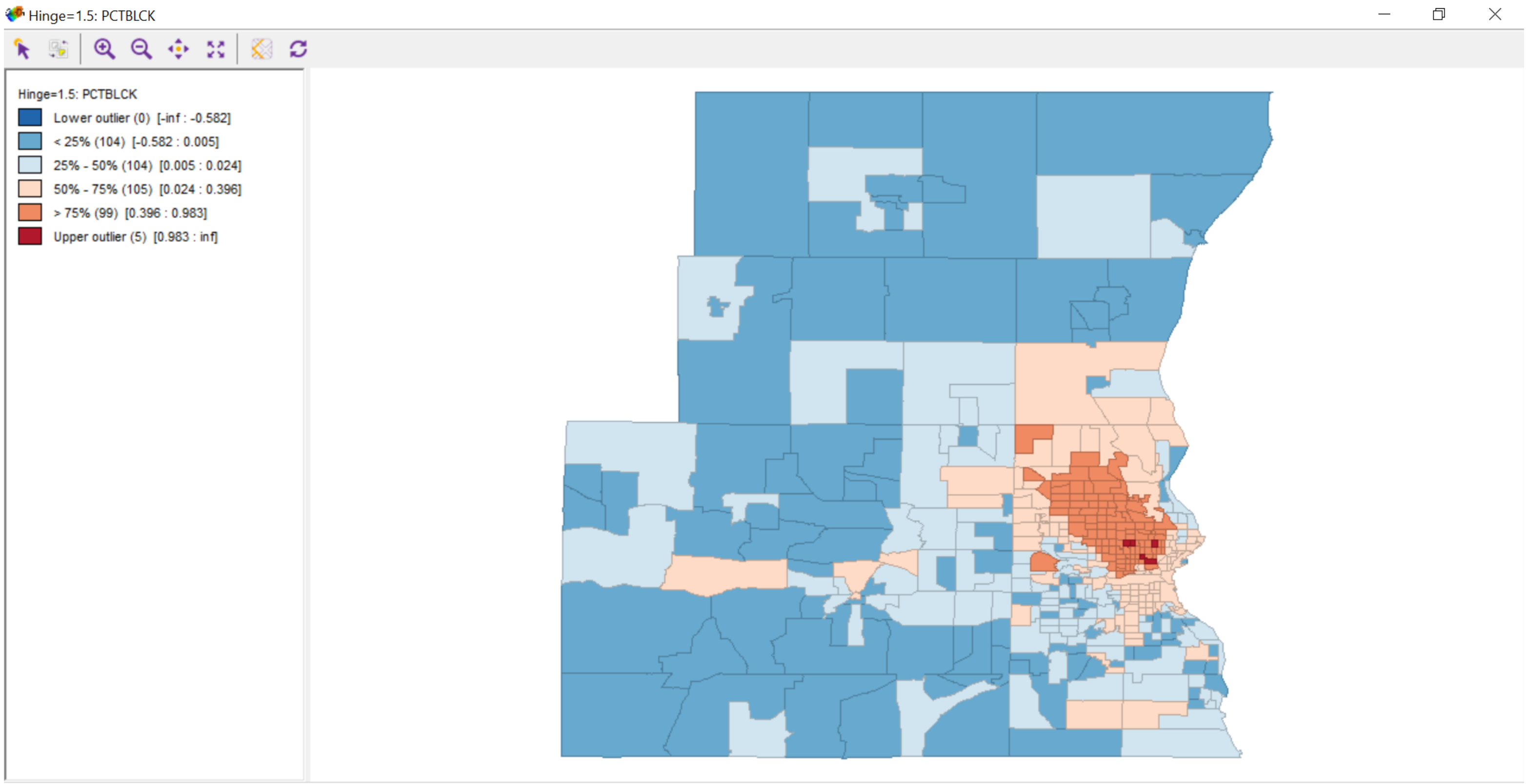
1. Go to *Space-Univariate LISA (or use icon)*
2. Select the spatial weights matrix milwaukee_queen.gal
3. Select the variable *PCTBLCK*
4. Select the cluster map option
5. Permutation options and higher significance levels are available through right-clicking

Moran scatterplot

1. Go to *Space-Univariate Moran (or use icon)*
2. Select the spatial weights matrix milwaukee_queen.gal
3. Select the variable *PCTBLCK*
4. Access the on-the-fly permutation visualization window through right-clicking

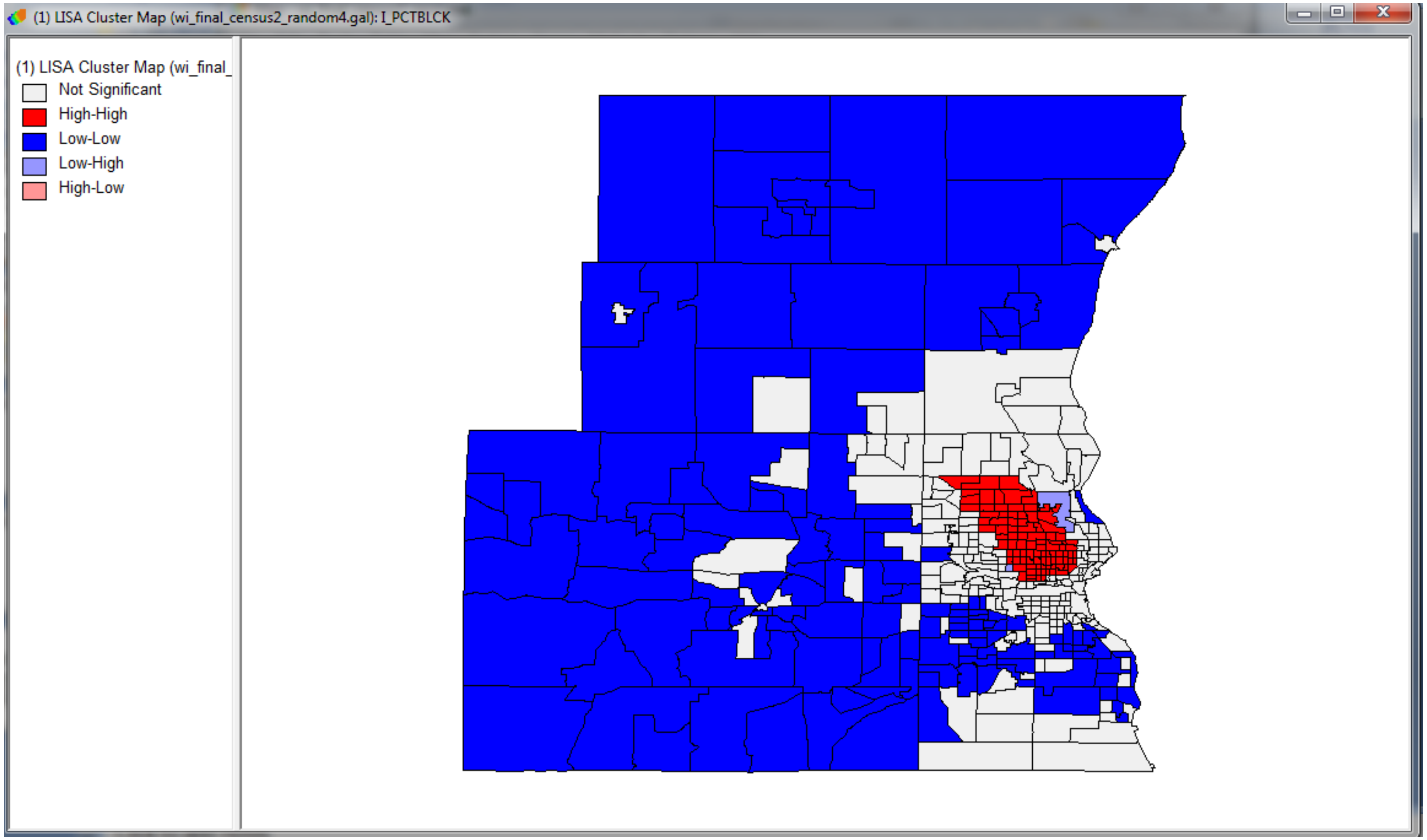
Spatial Autocorrelation – Example 2

Box Map



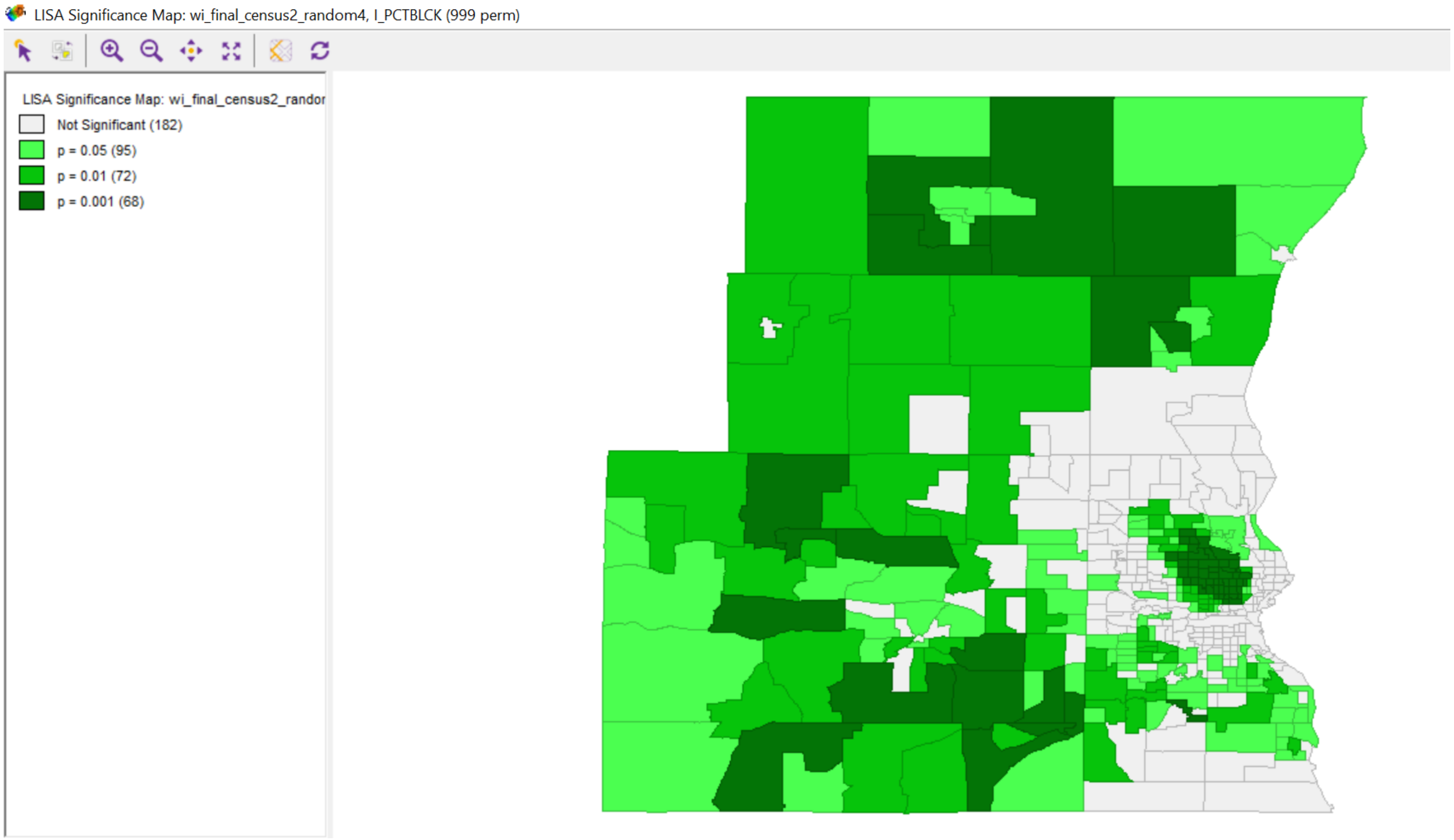
Spatial Autocorrelation – Example 2

LISA Map



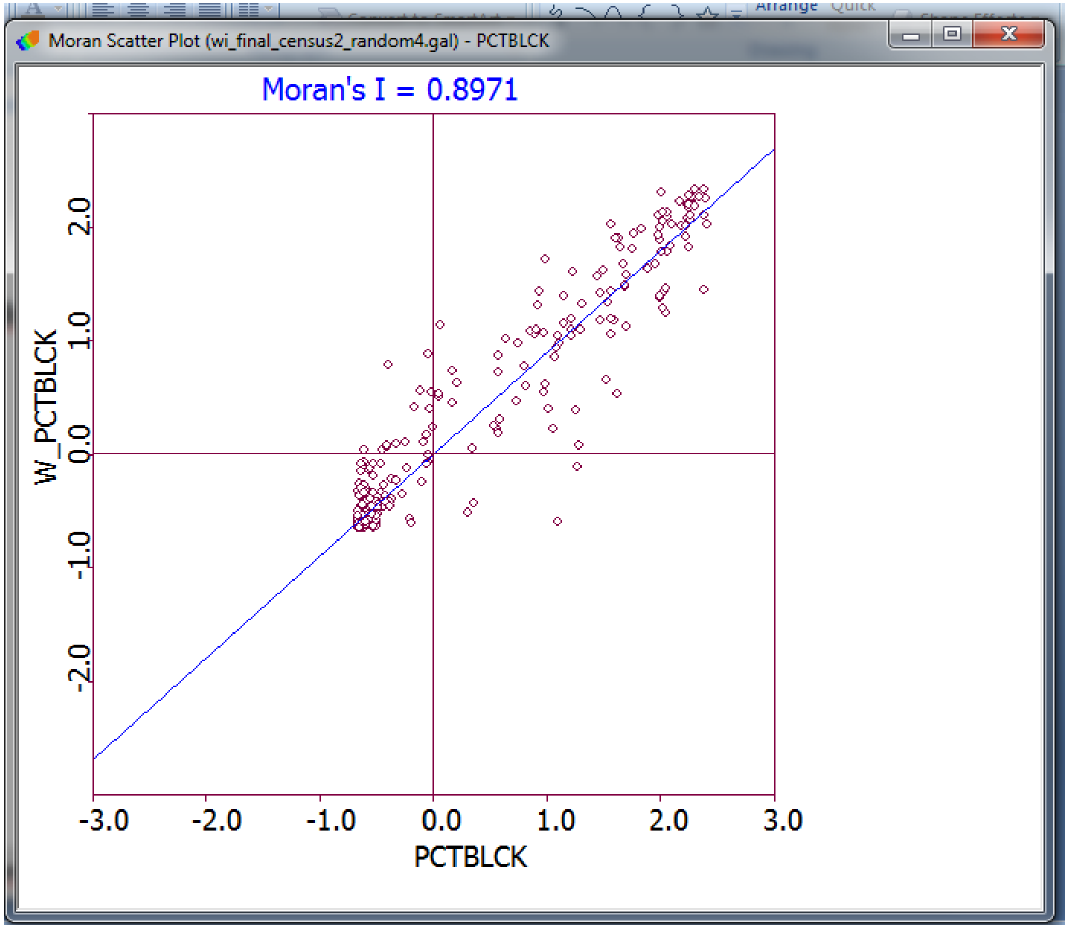
Spatial Autocorrelation – Example 2

Lisa significance Map



Spatial Autocorrelation – Example 2

Moran scatterplot



Spatial Regression

Significant spatial autocorrelation in the dependent variable or error term, can justify the incorporation of space in multiple linear regression models, giving rise to spatial regression models

$$Y = \alpha + \beta X + \varepsilon$$

$$Y = \alpha + \rho WY + \beta X + \varepsilon$$

$$Y = \alpha + \beta X + \varepsilon$$

$$\varepsilon = \lambda W \varepsilon$$

Spatial Regression

Basic modeling process:

1. Linear regression analysis $y = \alpha + \beta X + \varepsilon$
   - Classical model (linear regression)

2. Spatial regression analysis
   - 2.1. **Spatial Lag Model or Spatial Autoregressive Model (SAR)** $y = \alpha + \rho W y + \beta X + \varepsilon$
   - (autocorrelation in the dependent variable)

   - 2.2. **Spatial Error Model (SEM)** $Y = \alpha + \beta X + \varepsilon$
   - (autocorrelation in the error term)

$$\varepsilon = \lambda W \varepsilon$$

Spatial Regression

Other type of models

SDM – Spatial Durbin Model – autoregression also in the independent variables.

$$y = \alpha + \rho W y + \beta X + W X \theta + \varepsilon$$

Linear model

$$y = \boldsymbol{X}\beta + \mu$$

y ➢ Is the vector of values of the dependent variable and explained to the observations;

$\boldsymbol{X}$ ➢ is a matrix of values for the independent or explanatory variables for these observations;

β ➢ is a vector of coefficients and

μ ➢ is an error term

Linear model

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1k-1} \\ 1 & x_{21} & \dots & x_{2k-1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{k-1} x_{k-1i} + \varepsilon_i$$

Ordinary Least Square - OLS – 3 Basic conditions

1 - The error term must have zero mean and constant variance for all observations

$$E[\varepsilon_i] = 0 \; e \qquad E[\varepsilon_i{}^2] = \sigma^2$$

2 - The error values associated with observations *i* ad *j* should not be correlated:

$$E[\varepsilon_i \varepsilon_j] = 0 \qquad i \neq j$$

3 - The error term follows a normal distribution:

$$\varepsilon \sim (0, \sigma^2 I)$$

Spatial Regression

It is common that for variables with spatial distribution at least one of these last three conditions is not fulfilled

Therefore there is a bias on classical spatial regression equation

Process of estimation of spatial regression models

1. Developing statistical tests that allow:
   a) To check violation of the conditions of the classical model
   b) To check whether this is due to the existence of autocorrelation

2. If so……Model estimated incorporating the spatial structure of neighborhood relations (incorporation of the spatial relations matrix into the models).

Spatial Regression

Example: Auto-correlation for the dependent variable:

Vector of values: For each *yi*, calculate the weighted sum of the neighbor values ,

$$y = \rho \boldsymbol{W} y + \boldsymbol{X}\beta + \mu$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} = \rho \begin{bmatrix} w_{11} & \dots & w_{1n} \\ w_{21} & \dots & w_{2n} \\ \dots & \dots & \dots \\ w_{n1} & \dots & w_{nn} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{bmatrix} + \begin{bmatrix} 1 & x_{11} & \dots & x_{1k-1} \\ 1 & x_{21} & \dots & x_{2k-1} \\ \dots & \dots & \dots & \dots \\ 1 & x_{n1} & \dots & x_{nk-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \dots \\ \beta_{k-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \dots \\ \varepsilon_n \end{bmatrix}$$

Spatial Regression

Example: Auto-correlation for the dependent variable:

Nine territorial units (arranged contiguously) assigning each of them:

- A variable to be explained (or dependent) - *yi*
- An explanatory variable (or independent) - *xi*.

x_1, y_1	x_2, y_2	x_3, y_3
x_4, y_4	x_5, y_5	x_6, y_6
x_7, y_7	x_8, y_8	x_9, y_9

Spatial Regression

In the case of the territorial unit five (5), there is the following model development for the territorial gap, assuming a rook neighborhood matrix, standardized by line.

$$y_5 = \rho\left(\sum_j w_{5j} y_j\right) + x_{5k}\beta + \varepsilon_5$$

Spatial Regression

	Z1	Z2	Z3	Z4	Z5	Z6	Z7	Z8	Z9
Z1	0	0,5	0	0,5	0	0	0	0	0
Z2	0,33	0	0,33	0	0,33	0	0	0	0
Z3	0	0,5	0	0	0	0,5	0	0	0
Z4	0,33	0	0	0	0,33	0	0,33	0	0
Z5	0	0,25	0	0,25	0	0,25	0	0,25	0
Z6	0	0	0,33	0	0,33	0	0	0	0,33
Z7	0	0	0	0,5	0	0	0	0,5	0
Z8	0	0	0	0	0,33	0	0,33	0	0,33
Z9	0	0	0	0	0	0,5	0	0,5	0

$$Wy = \frac{\sum_{j=1}^{n} w_{ij} y_{ij}}{\sum_{j=1}^{n} w_{ij}}$$

$$y_5 = \rho(0{,}25 * y_2 + 0{,}25 * y_4 + 0{,}25 * y_6 + 0{,}25 * y_8) + x_5\beta + \varepsilon_5$$

x1, y1	x2, y2	x3, y3
x4, y4	x5, y5	x6, y6
x7, y7	x8, y8	x9, y9

Spatial Regression

$$y = \rho \boldsymbol{W_1} y + \boldsymbol{X}\beta + \mu \longrightarrow \mu = \lambda \boldsymbol{W_2} \mu + \varepsilon$$

- Assuming that the error term is auto correlated

y - vector of values for the dependent variable in each territorial unit;

β - vector of parameters associated with the independent variables

X - matrix of n values for each of the exogenous variables in each of the territorial units;

ε - vector of n error terms of normal distribution with a covariance matrix with covariance equal to $\sigma^2 I$ (constant) and zero average.

$$Var[\varepsilon_i] = \sigma^2 \boldsymbol{I} \qquad \varepsilon \sim N(0, \sigma^2 \boldsymbol{I}).$$

Spatial Regression

$$y = \rho \boldsymbol{W_1} y + \boldsymbol{X}\beta + \mu \longrightarrow \mu = \lambda \boldsymbol{W_2} \mu + \varepsilon$$

ρ - Autoregressive coefficient for the dependent variable – measures the average influence on the value of the variable *y* in each spatial unit of the same variable *y* in neighboring spatial units, i.e. indicates the proportion of the total variation of *y* explained by the spatial autocorrelation of *y*.

Spatial Regression

$$y = \rho \boldsymbol{W_1} y + \boldsymbol{X}\beta + \mu \longrightarrow \mu = \lambda \boldsymbol{W_2} \mu + \varepsilon$$

λ - Autoregressive coefficient for the error term

measures the average influence on the value of the variable y in each spatial unit of the error on the values for each neighboring spatial units, i.e. indicates the proportion of the total variation of y explained by the spatial autocorrelation of μ

Spatial Regression

$$y = \rho \boldsymbol{W}_1 y + \boldsymbol{X}\beta + \mu \quad \mu = \lambda \boldsymbol{W}_2 \mu + \varepsilon$$

$\boldsymbol{W}_1 y$ **spatial lag** – vector of values for the spatial auto correlated y variable

μ vector for the auto correlated elements in the error term

$$Var[\mu_i] = \sigma_i^2$$

or $E[\mu\mu'] = \boldsymbol{\Omega} = \sigma^2[(\boldsymbol{I} - \lambda\boldsymbol{W})'(\boldsymbol{I} - \lambda\boldsymbol{W})]^{-1}$

$\boldsymbol{W}_2 \mu$ **spatial error** - vector of values for the spatial auto correlated variable

Spatial Regression

Even when the classical regression results are satisfactory, the estimation of spatial models can provide a better approximation to the correlation between variables taking into account its location (and its neighbors) in the territory.

This estimate incorporating the territorial structure of neighborhood relations gives rise to the spatial autocorrelation through the matrix *W*.

Estimation process

- 1 - Evaluation of first-order effect
- 2 - Evaluation of second-order effects
- 3 - If there is autocorrelation estimate ... new model

Spatial Regression – First Order Effects

Phase 1 – Evaluation of first order effects

1.1. Classical regression analysis

1.2 Sensibility analysis for the existence of spatial autocorrelation (using a matrix of spatial structure)

Spatial Regression – First Order Effects

1.1. Regression analysis (classical model):

R2

Adjusted R2

Standard errors

F statistic

Maximum likelihood estimates - Criteria for Information

Number of conditional multicollinearity

Spatial Regression – First Order Effects

R2

Percentage of variation of the dependent variable explained by independent variables

For example, a value of 0.4 indicates that there is some relationship between the variables and it is reasonable to use a value estimated by the model.

The best model presents the best R2 (closest to a 1)

Spatial Regression – First Order Effects

R2

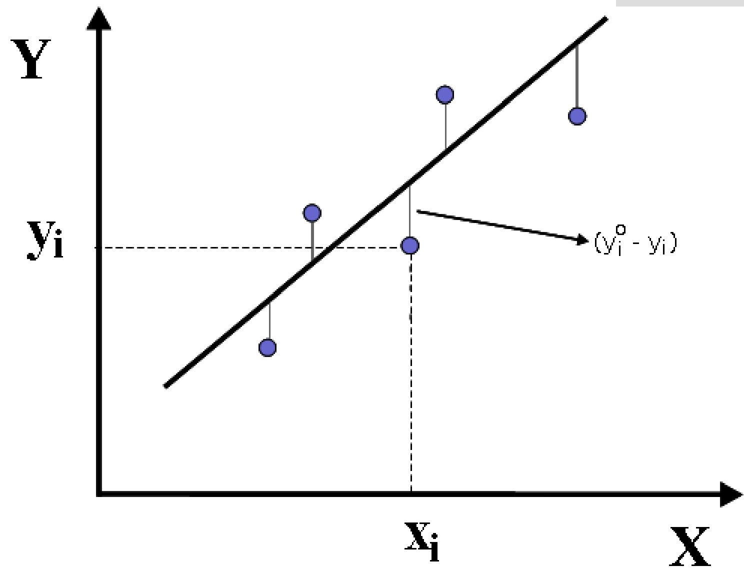
This measure is based on the decomposition of the total sum of squares of deviations from the mean for the dependent variable (TSS - total sum of squares) explained a sum of squares of deviations from predicted values (ESS - Explained sum of squares) and the sum of squared residuals (RSS - residual sum of squares).

$$\equiv \sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2 \equiv$$

$$\equiv 1 = \frac{ESS}{TSS} + \frac{RSS}{TSS} \equiv$$

$$\equiv R^2 = 1 - \frac{ESS}{TSS} = \frac{RSS}{TSS} \equiv$$

$$\equiv R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2}$$



Spatial Regression – First Order Effects

Adjusted R2

The adjusted R2 assesses changes in the number of variables involved, according to the following expression: where N represents the number of observations and K the number of variables.

$$R_a^2 = R^2 - \frac{(1 - R^2)(K - 1)}{N - K}$$

If this value does not increase (substantially) as variables are added, this means that adding another variable does not contribute to the model fit

Spatial Regression – First Order Effects

Standard errors

The estimated standard errors of regression coefficients and the probability of these coefficients are zero (null hypothesis) allows assessing their statistical significance.

$$H_0 \colon \beta_h = 0$$

Assuming that the regression residuals follow a normal distribution, significance is tested with a t-test, according to the expression:

$$t_h = \frac{\beta_h}{SE(\beta_h)}$$

Spatial Regression – First Order Effects

F statistic

The F statistic allows to test the joint hypothesis of all coefficients are nonzero.
For example, if R2 is 0.4 and the value of F test is 5.14, for a sample of 11 observations (Table F) and a significance level of 5% or p-value equal to 0.05 (which means we can reject the null hypothesis of independence between the dependent and independent variables with 95% probability) that means that likelihood of the relationship between variables is due to something more than chance.

Spatial Regression – First Order Effects

Maximum likelihood estimates (information criteria)

Alternative estimation method which admits
the classical regression bias.

Thus, for the adjustment capability of the models
that include spatial specification (estimated by maximum likelihood) we can compare it with the ability to adjust the classical regression models (estimated by the least squares method), This is to say one must construct similar statistics to make these capabilities comparable.

To compare **Classic Model** with **Spatial Model**

Spatial Regression – First Order Effects

Maximum likelihood estimates (information criteria)

The logarithm of this function (or log likelihood L) is a measure of adjustment alternative to classical regression R2 and comparable with similar measures of territorial adjustment regression.

The higher the L the better the model

- From the value of L are built information criteria, which consider a penalty based on the number of degrees of freedom. The general expression is:

$$CI = -2L + f(K, N)$$

Spatial Regression – First Order Effects

Maximum likelihood estimates (information criteria)

Akaike information criterion and $$CIAkaike = -2L + 2K$$

Schwarz information criterion $$CISwartz = -2L + Kln(N)$$

The smaller the values of information criteria of *Akaike* and *Swartz*, the better the model.

Spatial Regression – First Order Effects

Conditional multicollinearity number

- Measure of the degree of dependence between the explanatory variables, and hence its ability to provide independent information on the dependent variable.
- If there is a strong correlation between the explanatory variables, then there is a poor initial specification of the model.

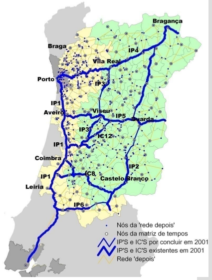
There is no multicollinearity when this number is less than 30

Spatial Regression – First Order Effects Example

$$TP_{9101} = \alpha + \beta_1 TP_{8191} + \beta_2 TD_{91} + \beta_3 ESE_{91} + \beta_4 TAD^{60}{}_{9101} + \varepsilon$$

TP9101 - Rate of population trends between 1991 and 2001

TP8191 - Rate of population trends between 1981 and 1991



TD91 - Unemployment rate in 1991

ESE1991 - Proportion of people with 12º grade of education

TAD - Rate of evolution of the population daily accessible to less than 60 minutes between 1991 and 2001

Spatial Regression – First Order Effects

1.1. Regression analysis (classical model):

R2

Adjusted R2

Standard errors

F statistic

Maximum likelihood estimates - Criteria for Information

Number of conditional multicollinearity

Spatial Regression – First Order Effects Example

REGRESSION/SUMMARY OF OUTPUT: ORDINARY LEAST SQUARES ESTIMATION

Dependent Variable : **TP9101** Number of Observations: 86
Mean dependent var. :-6.60891 Number of Variables: 5
S.D. dependent var. : 7.92788 Degrees of Freedom: 81
R-squared : 0.359590 / Adjusted R-squared : 0.327965
F-statistic: 11.3704 / Prob. (F-statistic) :2.25769e-007
Log likelihood: -280.919
Akaike info criterion: 571.838
Schwarz criterion: 584.11

Variable	Coefficient	Std.Error	t-Statistic	Probability
CONSTANT	-13.98978	3.019991	-4.632393	0.0000136
TP8191	0.1623603	0.08670297	1.872604	0.0647312
ESE91P	2.514192	0.4944105	5.085231	0.0000023
TD91	-0.7593184	0.3826478	-1.98438	0.0505998
TAD60M	0.01376335	0.01031015	1.334932	0.1856386

Multicollinearity Condition Number 10.26814

Spatial Regression – First Order Effects

1.2 Sensibility analysis for the existence of spatial autocorrelation

Terms of estimation by the method of least squares

Normality of residuals

Heterosckedasticity of the residuals

Spatial Regression – First Order Effects

Remembering: Terms of estimation by the method of least squares

The random error terms follow a normal distribution. $N \sim \left(0, \sigma^2\right)$

The random error has mean zero (there is no bias in the regression equation).

$$E\left[\varepsilon_i\right] = 0$$

The random error terms are uncorrelated and have constant variance (It is the property of the 'homoskedasticity')

$$\sigma^2(\varepsilon_i) = \sigma^2$$

Spatial Regression – First Order Effects

Normality of residuals

Checking normality of residuals, the null hypothesis:

$$H_0 : \varepsilon \sim (0, \sigma^2 \boldsymbol{I})$$

Jarque-Bera test,

where S represents a measure of symmetry (skeweness) and K represents a measure of flatness or kurtosis, which measure the departure from a normal distribution (symmetric and bell-shaped).

$$JB = \frac{n}{6}\left(S^2 + \frac{(K-3)^2}{4}\right)$$

Ideal: Do not reject the null hypothesis (**p-values higher**) - there is a normal distribution in the residuals (necessary for other tests)

Spatial Regression – First Order Effects

Heteroscedasticity in the residuals

If the regression residuals do not have a constant variance over the territory there is no hypothesis of homoscedasticity.

As a result, the values of confidence intervals and test application of the method of least squares are not reliable. There are several types of tests but all start from the hypothesis of homoscedasticity :

$$H_0 : E[\varepsilon \varepsilon'] = \sigma^2 \boldsymbol{I}$$

Testing for heterokedasticity

Breuch-Pagan, Koenker-Basset and White.

If the null hypothesis is rejected (p-values higher). There is heterokedasticity.

Spatial Regression – First Order Effects

The presence of heterokedasticity can mean the existence of spatial autocorrelation, which is confirmed (or not) in testing for the effects of 2nd order.

- When heterokedasticity and autocorrelation are present simultaneously, the first is very dependent on the second.

This means that if the heterokedasticity is due only to the existence of spatial autocorrelation, the problem of model specification revealed by the heterokedasticity can be solved with the incorporation of autocorrelation into the model.

Spatial Regression – First Order Effects

TEST ON NORMALITY OF ERRORS

TEST	DF	VALUE	PROB
Jarque-Bera	2	3.613382	0.1641966

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	8.90696	0.0634676
Koenker-Bassett test	4	6.183338	0.1858685

SPECIFICATION ROBUST TEST

TEST	DF	VALUE	PROB
White	14	12.45323	0.5699537

Spatial Regression – Second Order Effects

Degree and type of spatial auto regression

Tests that require specific programs (Like GeoDa), which allow the inclusion of territorial structure in the form of a neighborhood matrix W. The classical model can be made robust with the explicit consideration of spatial effects

Null hypothesis: lack of auto correlation

$$H_0: \rho = 0$$

$$H_0: \lambda = 0$$

Spatial Regression – Second Order Effects

Degree of spatial autocorrelation - Moran's I

Indicator of global autocorrelation, applied to the regression residuals.

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

Spatial Regression – Second Order Effects

Type of spatial autocorrelation - Lagrange multipliers

Statistics based on maximum
likelihood function including robust versions:

(Lagrange Multiplier Lag - or LML)
Lagrange multiplier for the model of spatial lag

Lagrange multiplier for the model error territorial
(Lagrange Multiplier Error - or LME)

Spatial Regression – Second Order Effects

Type of spatial autocorrelation - Lagrange multipliers

(LMLR) - The robust version for lag model tests the possibility of territorial spatial lag in the presence of autocorrelation in the error term and

(LMER) - The robust version for the error model tests the possibility of spatial error in the presence of a dependent variable auto correlated

Spatial Regression – Second Order Effects

Type of spatial autocorrelation - Lagrange multipliers

Misspecification of the model:
Statistics associated with testing and LML, LME significant, but those associated with its robust tests are not significant.

Desirable: One type of territorial self-correlation is prevalent through the statistics LML, LME, and that one of the statistics is no longer significant the statistics LMLR, LMER value, since the estimation process becomes simpler (means there is only one type of spatial regression in the model).

Spatial Regression – Second Order Effects

If the correlation coefficient is significant in the spatial lag model, this means that the power of independent variables to explain the dependent variable is largely due to the values of this variable in neighboring territorial units, and

If the correlation coefficient is significant in the spatial error model, that means there is still some auto-correlation in the dependent variable, but due to variables that are still missing in the model. Therefore, the autocorrelation is detected in the error term in the regression.

Spatial Regression – Second Order Effects

Degree and type of territorial self-correlation - Lagrange multipliers

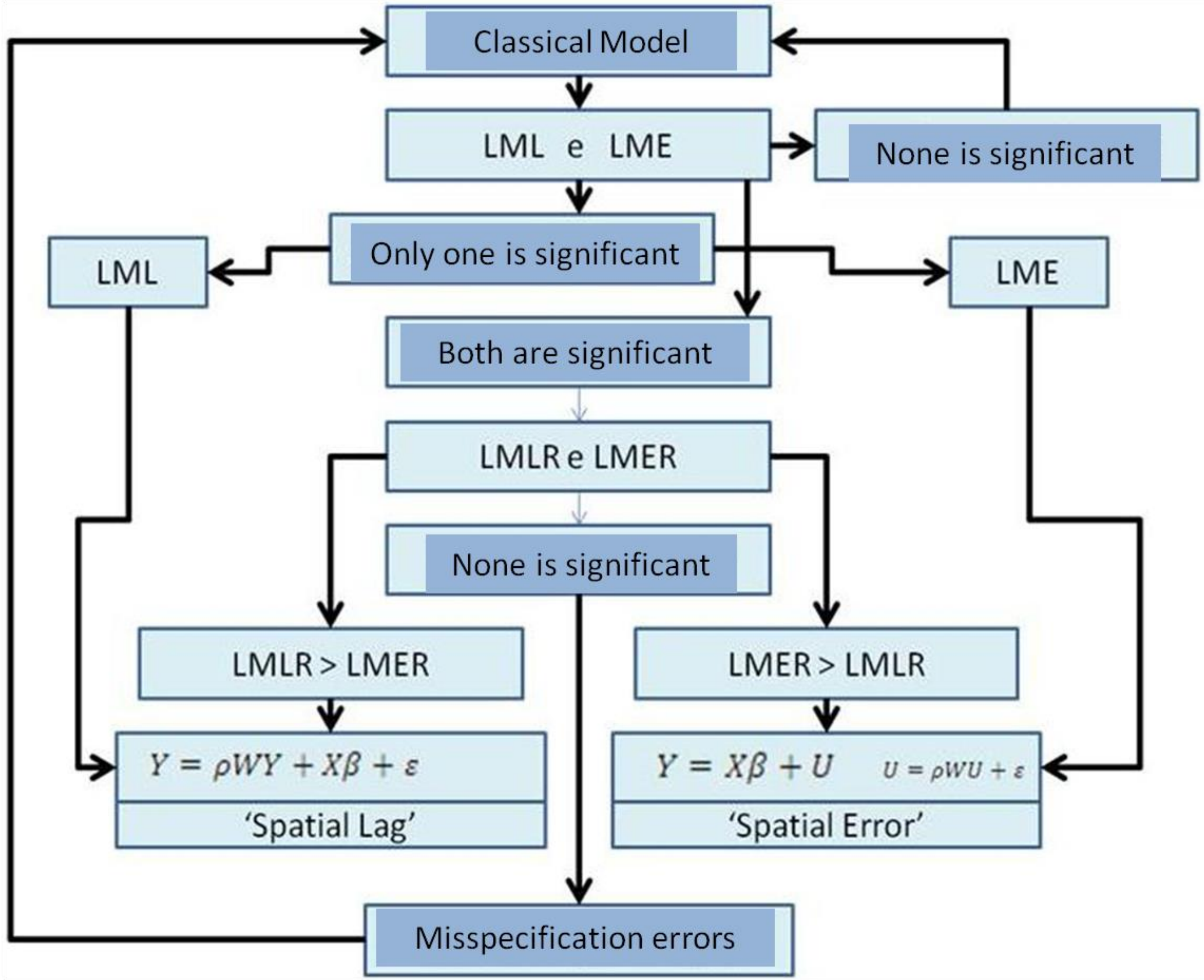
Moran's I is significant (value of 0.26 for I and 99% probability of the rejection of the hypothesis that there is no territorial self correlation with a p-value of 0.0000339).

Both types of auto-correlation are present (lag and error territorial) mainly in the error case. In robust statistics, the spatial lag is no longer significant in the presence of a spatial error autocorrelation

DIAGNOSTICS FOR SPATIAL DEPENDENCE

TEST	MI/DF	VALUE	PROB
Moran's I (error)	**0.263852**	4.14541	0.0000339
Lagrange Multiplier (lag)	1	8.8163366	0.0029855
Robust LM (lag)	1	0.1561955	0.6926837
Lagrange Multiplier (error)	1	13.7050658	0.0002139
Robust LM (error)	1	5.0449247	0.0246982

Spatial Regression – Estimation process



Spatial Regression - Estimation process

When appropriate to estimate the spatial model recommended by previous tests, estimate and perform new tests (***Breuch- Pagan and Likelihood Ratio***)

Breuch-Pagan (BP) - If the test statistic associated presents an high value and it is significant, then heteroskedasticity remains a property of the error variance.

Likelihood Ratio (LR) – Test associated with the correspondent Lagrange multiplier. If the value of the statistic associated with this test is significant, then the consideration of the spatial specification is correct.

Spatial Regression - Estimation process

Possible specification errors:

- The relationship between the explanatory variables and the dependent variable is nonlinear;
- The variables are not adequate to explain the process in question;
- The matrix of territorial structure is not adequate to explain spatial relations for that particular model

Spatial Regression - Estimation process

Solutions for specification errors:

Reformulating the model

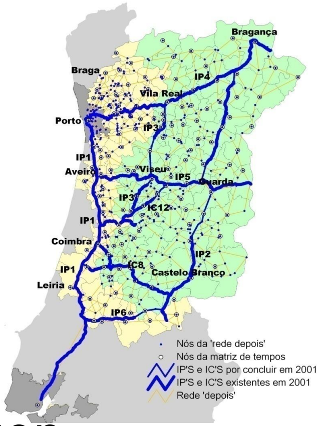
- Using another type of regression models
- Variable transformation/other variables
- Another neighborhood matrix

Spatial Regression - Estimation process example

$$TP_{9101} = \alpha + \beta_1 TP_{8191} + \beta_2 TD_{91} + \beta_3 ESE_{91} + \beta_4 TAD^{60}{}_{9101} + \varepsilon$$

TP9101 - Population evolution between 1991 and 2001

TP8191 - Population evolution between 1981 and 1991



TD1991 – Unemployment rate1991

ESE1991 – Proportion of people with 12º grade of education

TAD - Rate of evolution of the population daily accessible to less than 60 minutes between 1991 and 2001

Spatial Regression - Estimation process example

Spatial Lag Model

SPATIAL LAG MODEL - MAXIMUM LIKELIHOOD ESTIMATION

Dependent Variable :	**TP9101**	Number of Observations:	86
Mean dependent var. :	-6.60891	Number of Variables :	6
S.D. dependent var. :	7.92788	Degrees of Freedom :	80
Lag coeff. () :	0.378754		
R-squared :	0.439391		
		Log likelihood:	**-276.612**
Sq. Correlation : -		**Akaike info criterion:**	**565.224**
Sigma-square :	35.235	**Schwarz criterion:**	**579.95**
S.E of regression :	5.9359		

Variable	Coefficient	Std.Error	z-value	Probability
W_TP9101	0.3787538	0.1145022	3.30783	0.0009403
CONSTANT	-13.57934	2.845178	-4.772756	0.0000018
TP8191	0.1051584	0.07908267	1.329728	0.1836082
ESE91P	2.757667	0.4494318	6.135896	0.0000000
TD91	-0.7200998	0.347458	-2.07248	0.0382206
TAD60	0.01187408	0.009362363	1.268278	0.2046990

Spatial Regression - Estimation process example

Spatial Error Model

SPATIAL ERROR MODEL - MAXIMUM LIKELIHOOD ESTIMATION

```
Dependent Variable    :       TP9101  Number of Observations:   86
Mean dependent var    :    -6.608910  Number of Variables   :    5
S.D. dependent var    :     7.927877  Degree of Freedom     :   81
Lag coeff. (Lambda)   :     0.602106

R-squared             :     0.520541  R-squared (BUSE)      : -
Sq. Correlation       : -                  Log likelihood          : -272.490411
Sigma-square          :    30.134607       Akaike info criterion :     554.981
S.E of regression     :       5.4895       Schwarz criterion       :  567.252559
```

Variable	Coefficient	Std.Error	z-value	Probability	
CONSTANT	-18.87322	2.839847	-6.645857	0.0000000	
TP8191	0.04972097	0.07255608	0.6852765	0.4931693	
ESE91P	3.198523	0.4122815	7.758105	0.0000000	
TD91	-0.6516037	0.3393885	-1.919934	0.0548661	
TAD60	0.004910706	0.009846709	0.4987155	0.6179798	
LAMBDA	0.6021061	0.1032605	5.830943	0.0000000	0.0000000

Spatial Regression - Estimation process example

Spatial Lag - The statistics associated with the test for heteroskedasticity indicates that this is still present although to a significance level of 10% (90% chance of rejecting the hypothesis of homoskedasticity with p value of 0.06).

The test of spatial autocorrelation confirms the validity of the model (p-value of 0.003).

REGRESSION DIAGNOSTICS

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	DF	VALUE	PROB
Breusch-Pagan test	4	9.011116	0.0608222

DIAGNOSTICS FOR SPATIAL DEPENDENCE

SPATIAL LAG DEPENDENCE FOR WEIGHT MATRIX : **rook5.GAL**

TEST	DF	VALUE	PROB
Likelihood Ratio Test	1	8.613813	0.0033362

Spatial Regression - Estimation process example

Spatial Error - The test indicates that the heteroskedasticity dropped out of the model. It is no longer significant the existence of heterokedasticity

The test of spatial autocorrelation confirms the validity of the model.

REGRESSION DIAGNOSTICS

DIAGNOSTICS FOR HETEROSKEDASTICITY

RANDOM COEFFICIENTS

TEST	**DF**	**VALUE**	**PROB**
Breusch-Pagan test	4	7.493963	0.1119758

DIAGNOSTICS FOR SPATIAL DEPENDENCE

SPATIAL ERROR DEPENDENCE FOR WEIGHT MATRIX : **rook5.GAL**

TEST	**DF**	**VALUE**	**PROB**
Likelihood Ratio Test	1	16.8573	0.0000403

Spatial Regression - Estimation process example

Comparing values

Criteria	Models		
	Classic Model	Spatial Lag Model	Spatial Error Model
Log Likelihood	-280,919	-276,612	-272,49
Akaike	571,839	565,224	554,981
Schwarz	584,11	579,95	567,252
R2	0,36		
ρ		0,38***	
λ			0,60***
Constant	-13,99***	-13,58***	-18,87***
TP8191	0,16*	0,11	0,049
ESE91	2,5***	2,75***	3,20***
TD91	-0,76**	-0,72**	-0,65*
TAD60M	0,014	0,011	0,0050
Heterokedasticity		9,01*	7,5
Spatial dependence		8,61***	16,85***

Spatial Regression - maps

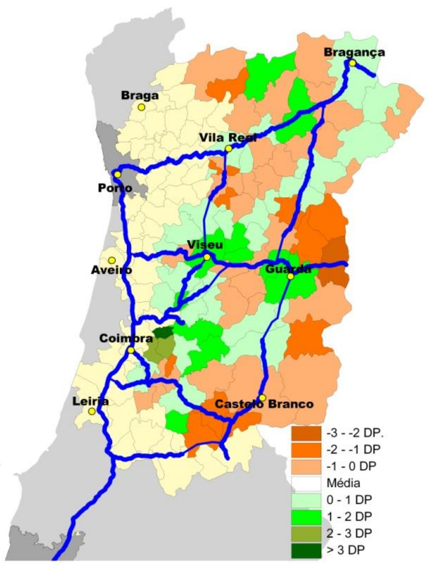
The graphic expression of some of the results of estimation of the models is useful for identifying which types of territorial relations determined the statistical significance of spatial autocorrelation

- **Maps of residuals and**
- **Maps of estimated values.**

Spatial Regression - maps

Classic model - Map of residuals,

- It identifies many outliers, corresponding to residuals with values greater than the expected value (the dark green, values underestimated by the model) and with values below the expected value (or orange brown, overestimatec model).



$$TP_{9101} = \alpha + \beta_1 TP_{8191} + \beta_2 TD_{91} + \beta_3 ESE_{91} + \beta_4 TAD^{60} + \varepsilon$$

Spatial Regression - maps

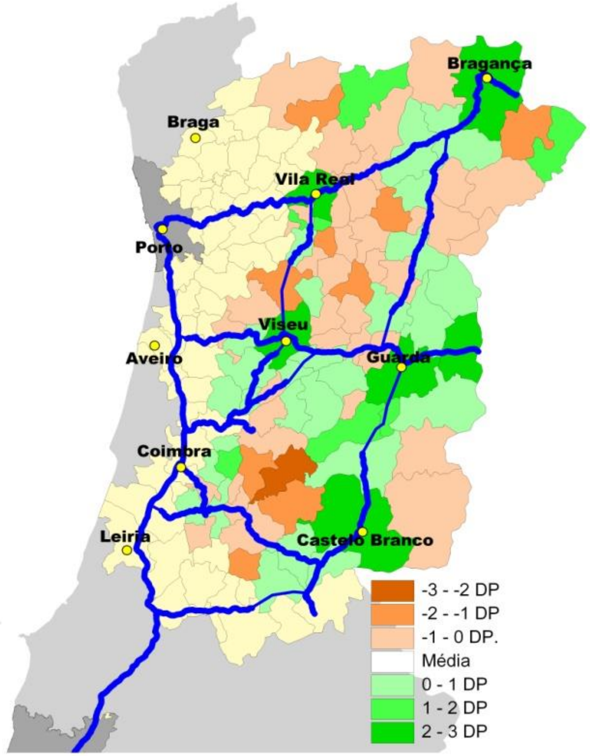
Error model - Map of estimated values

On the map of estimated values for the model error, it appears that there are values well above average and well below the average values (more than two standard deviations). On the other hand, this map also identifies groups of contiguous spatial units with similar characteristics. The model therefore identifies a tendency for there is variability in the relationship modeled, depending on the relative position of the units.

$$TP_{9101} = \alpha + \beta_1 TP_{8191} + \beta_2 TD_{91} + \beta_3 ESE_{91} + \beta_4 TAD^{60} + (I - \lambda W)^{-1} \varepsilon$$

Spatial Regression - maps

Error model - Map of estimated values



Spatial Regression - Aplications

The combination of the programs:

ARCGIS or QGIS (to better operate with data transformations)

GEODA (to operate simple Spatial analysis, in a exploratory basis and searching for relations between transport variables and development variables)

MATLAB (to develop more complex spatial analysis)

R (idem)

Opening new fields in research which results that can improve drastically knowledge and policy advice for a better articulation between Transport Systems Planning and Development or Urban Planning.

References

Le Sage & Pace Kelly (2009). Spatial Econometrics. CRC Press

Getis, Mur and Zoller (2004). Spatial Econometrics and Spatial Statistics. Ed. Palgrave.

Anselin, L. (1988). Spatial Econometrics: Methods and Models. Boston, Massachusetts: Kluwer Academic Publishers.

Anselin, L. (2003). GeoDa 0.9 User's Guide. Spatial Analysis Laboratory, University of Illinois, Urbana-Champaign, IL, pp. 88-105.

Anselin, Luc. (2003). An Introduction to Spatial Autocorrelation Analysis with GeoDa.

LeSage, J. (2009), Introduction to Spatial Econometrics, CRC Press.

INSTITUTO SUPERIOR TÉCNICO

FEUP

Phd Program in Transportation

Transport Demand Modeling

Anabela Ribeiro - 2021

anabela@dec.uc.pt

Spatial Regression Models