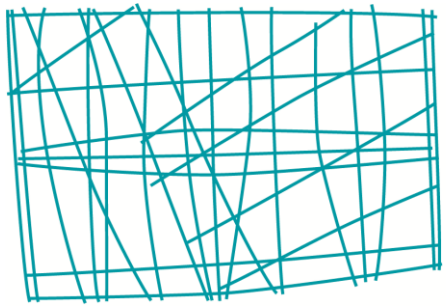


NEREUS

Núcleo de Economia Regional e Urbana
da Universidade de São Paulo

The University of São Paulo
Regional and Urban Economics Lab

Lecture 3: Exploratory Spatial Data Analysis (ESDA)

Prof. Eduardo A. Haddad

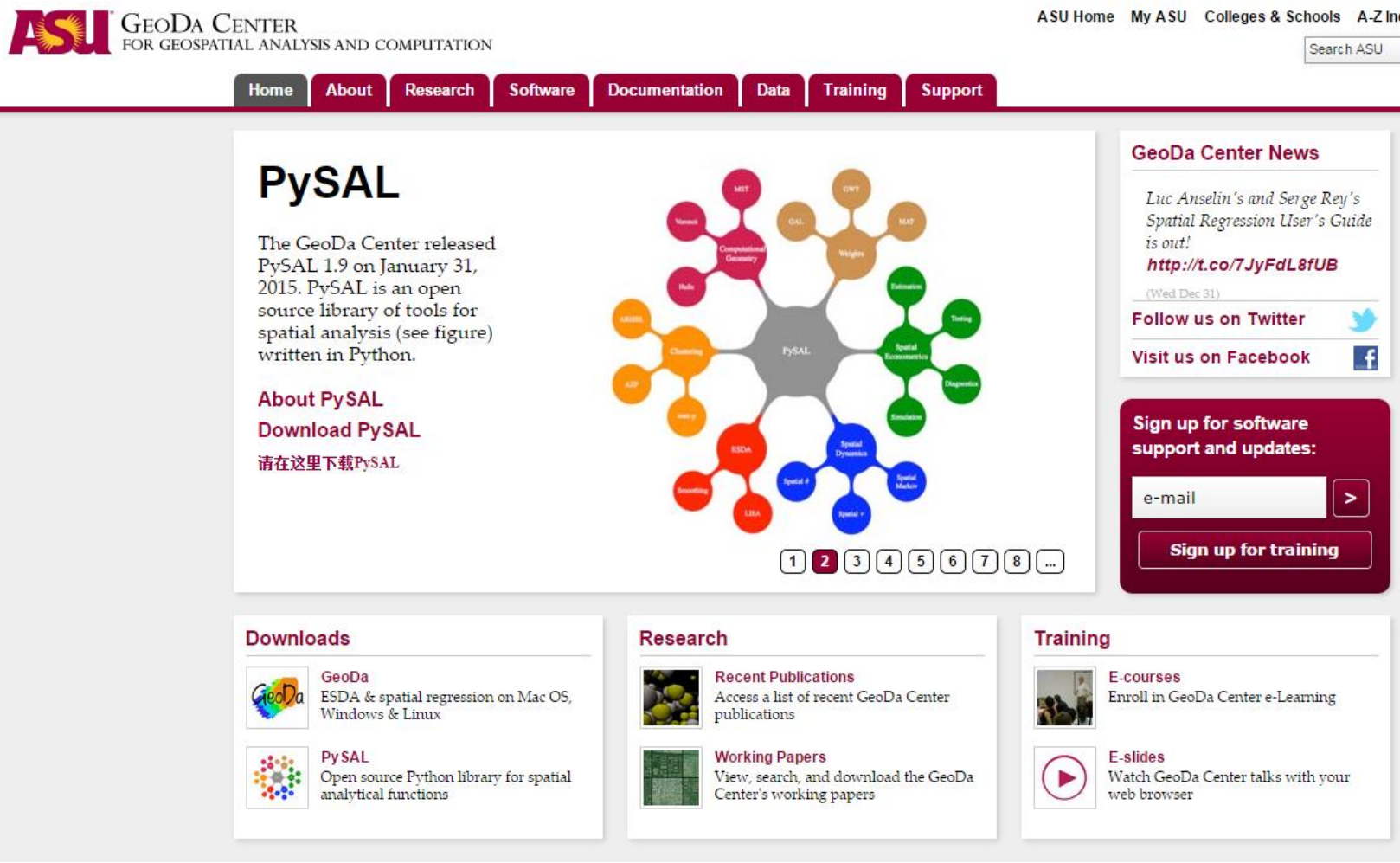
Key message

Spatial dependence

First Law of Geography (Waldo Tobler):

"Everything is related to everything else, but near things are more related than distant things"

GeoDa – https://geodacenter.asu.edu/



Spatial analysis

Locational invariance

- Spatial analysis is **not** locationally invariant
- The results change when the locations of the study objects change
- Where matters!

Mapping and Geovisualization

- Showing interesting patterns

Exploratory Spatial Data Analysis

- Discovering interesting patterns

Spatial Modeling

- Explaining interesting patterns

ESDA

Exploratory data analysis (EDA)uses a set of techniques to:

- maximize insight into a data set
- uncover underlying structures
- extract important variables
- detect outliers and anomalies
- test underlying assumptions
- suggest hypotheses
- develop parsimonious models

E**S**DA includes spatial attributes of the data

Geovisualization

Beyond mapping:

- Combining map and scientific visualization methods
- Exploit human pattern recognition capabilities

Statistical maps

- Innovative map devices (quantile map, percentile map, box map, standard deviation map)

Map movie

Exercise 1 (mapping)

Open and close a project

Load a shape file with the proper indicator (Key)

Select functions from the menu or toolbar

Make a simple choropleth map

Select items in the map

Data: Morocco (gdppc_07), Brazil_UF (Y00, G00)

Exercise 2 (data)

Open and navigate the data table

Select and sort items in the table

Create new variables in the table

Data: Morocco (gdppc_03, gdppc_07)

Which regions presented the highest per capita GRP growth from 2003 to 2007 in Morocco?

Table > Field calculation

Exercise 3 (EDA)

This exercise illustrates some basic techniques for exploratory data analysis, or EDA. It covers the visualization of the non-spatial distribution of data by means of a histogram and box plot, and highlights the notion of linking, which is fundamental in GeoDa.

Data: Brazil_UF (Y00, G00)

Histogram:

- The histogram is a discrete approximation to the density function of a random variable

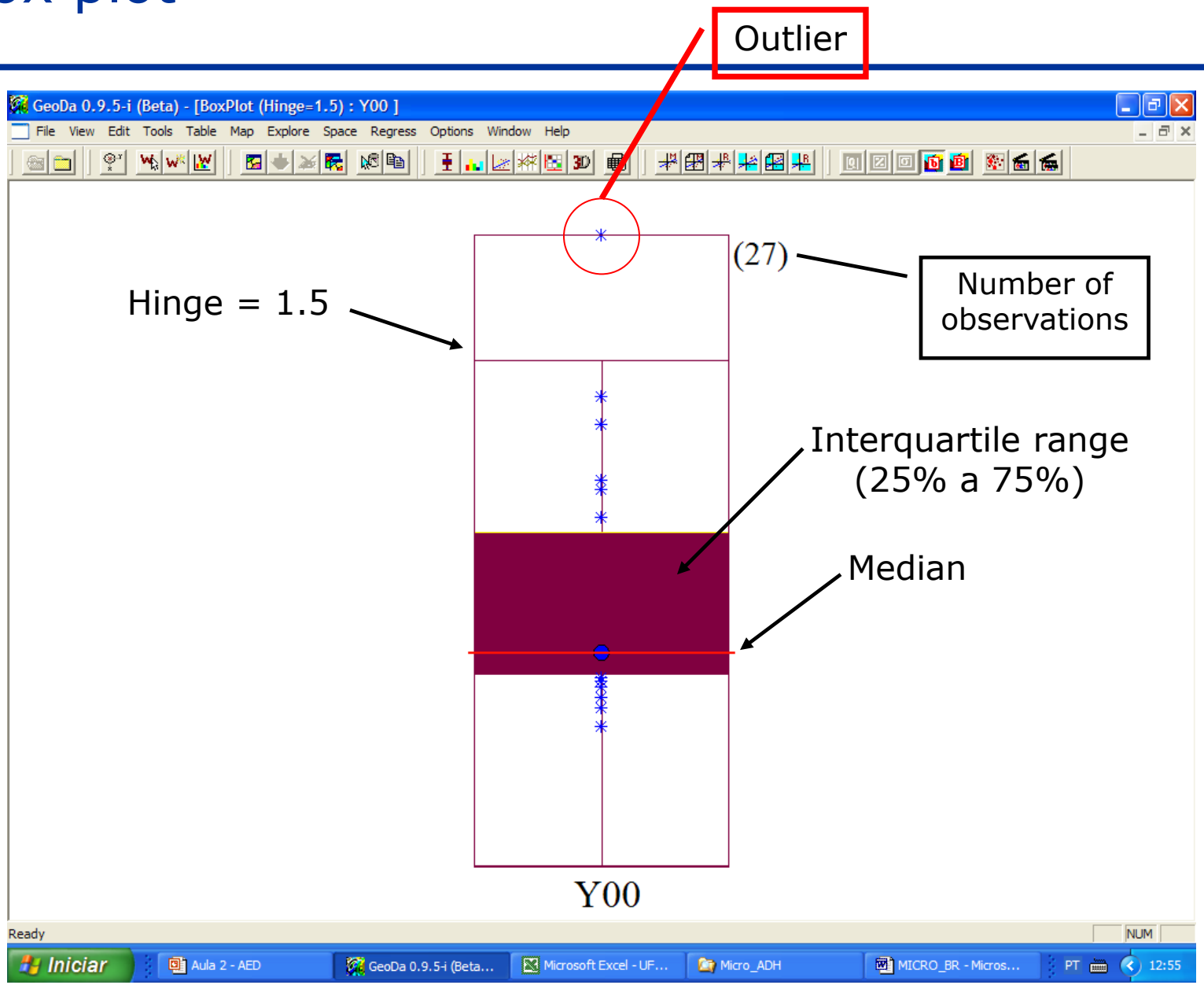
Exercise 3 (EDA)

Box plot:

- It shows the median, first and third quartile of a distribution (the 50%, 25% and 75% points in the cumulative distribution) as well as a notion of outlier.

- An observation is classified as an **outlier** when it lies more than a given multiple of the interquartile range (the difference in value between the 75% and 25% observation) above or below respectively the value for the 75^{th} percentile and 25th percentile. The standard multiples used are 1.5 and 3 times the interquartile range.

Box plot


Outlier
GeoDa 0.9.5-i (Beta) - [BoxPlot (Hinge=1.5) : Y00]
File View Edit Tools Table Map Explore Space Regress Options Window Help
(27)
Number of observations
Hinge = 1.5
Interquartile range (25% a 75%)
Median
Y00
Ready
NUM
Iniciar
Aula 2 - AED
GeoDa 0.9.5-i (Beta...
Microsoft Excel - UF...
Micro_ADH
MICRO_BR - Micros...
PT
12:55

Exercise 3 (EDA)

Scatter plot:

- The the visualization of the bivariate association between variables can be done by means of a scatter plot
- Converting the scatter plot to a correlation plot, in which the regression slope corresponds to the correlation between the two variables (as opposed to a bivariate regression slope in the default case).
- Any observations beyond the value of 2 can be informally designated as outliers

Exercise 4 (spatial scale and rate of density)

Inference can change with scale

Problem of spatial aggregation

- State v. Micro-region

Intensity maps

- Extensive variable tends to be correlated with
- size (such as area or total population)
- Rate or density is more suitable for a choropleth map, and is referred to as a intensive variable.

Data: Brazil_UF (Y00) and Brazil_MR (RENDAPC, RENDA, POP)

Spatial dependence

What happens at one place depends on events in nearby places

All things are related but nearby things are more related than distant things (Tobler)

Categorizing

- Type: substantive *versus* nuisance
- Direction: positive *versus* negative

Issues

- Time *versus* space
- Inference

Concepts

Spatial weights

Spatial lag

Spatial autocorrelation

Spatial weights matrix

Definition:

- N by N positive matrix W, elements w_{ij}
- w_{ij} nonzero for neighbors, 0 otherwise
- $w_{ii} = 0$, no self-neighbors

Geographic weights (contiguity, distance, general, graph-based weights)

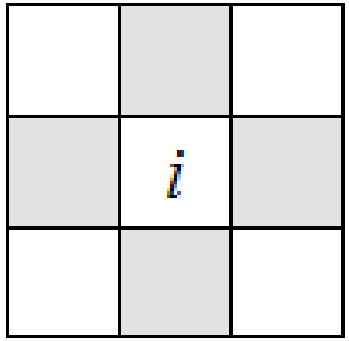
Socio-economic weights

Contiguity weights

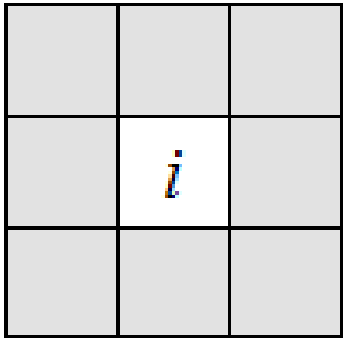
Contiguity: sharing a common boundary of non-zero length

Three views of contiguity:

- rook
- bishop
- queen



rook



queen

Exercise 5 (spatial weights)

Contiguity-based spatial weights

Create a first order contiguity spatial weights file from a polygon shape file, using both rook and queen criteria (.GAL)

Connectivity structure of the weights in a histogram

Higher order contiguity

Data: Brazil_UF

Exercise 6 (spatial weights)

Distance-based spatial weights

Create a distance-based spatial weights file from a point shape file, by specifying a distance band

Adjust the critical distance

Create a spatial weights file based on a k-nearest neighbor criterion

Data: Hedonic; Brazil_UF

Spatially lagged variables

Spatially lagged variables are an essential part of the computation of spatial autocorrelation tests and the specification of spatial regression models

Wy: is the average of a variable in the neighboring spatial units (e.g. WY00 is the average per capita income of the neighbors)

Exercise 7 (spatial lag)

Scatter plot of Y00 and WY00

Use standardized values

Data: Brazil_UF (Y00)

Clustering

Global characteristic

Property of overall pattern = all observations are like values more grouped in space than random

Test by means of a global spatial autocorrelation statistic

No location of the clusters determined

Clusters

Local characteristic

Where are the like values more grouped in space than random?

Property of local pattern = location-specific

Test by means of a local spatial autocorrelation statistic

Local clusters may be compatible with global spatial randomness

Spatial autocorrelation statistic

Formal test of match between value similarity and locational similarity

Statistic summarizes both aspects

Significance: how likely is it (p-value) that the computed statistic would take this (extreme) value in a spatially random pattern?

Attribute similarity

Summary of the similarity or dissimilarity of a variable at different locations:

- Variable y at locations i, j with $i \neq j$

Measures of similarity:

- Cross product: $y_i y_j$

Measures of dissimilarity:

- Squared differences: $(y_i - y_j)^2$
- Absolute differences: $|y_i - y_j|$

Global spatial autocorrelation (Moran's I)

$$I = \left(\frac{n}{S_0}\right) \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} z_i z_j}{\sum_{i=1}^{n} z_i^2}$$

$z_i = x_i - \bar{x}$

Cross-product statistic

Similar to a correlation coefficient

Value depends on W

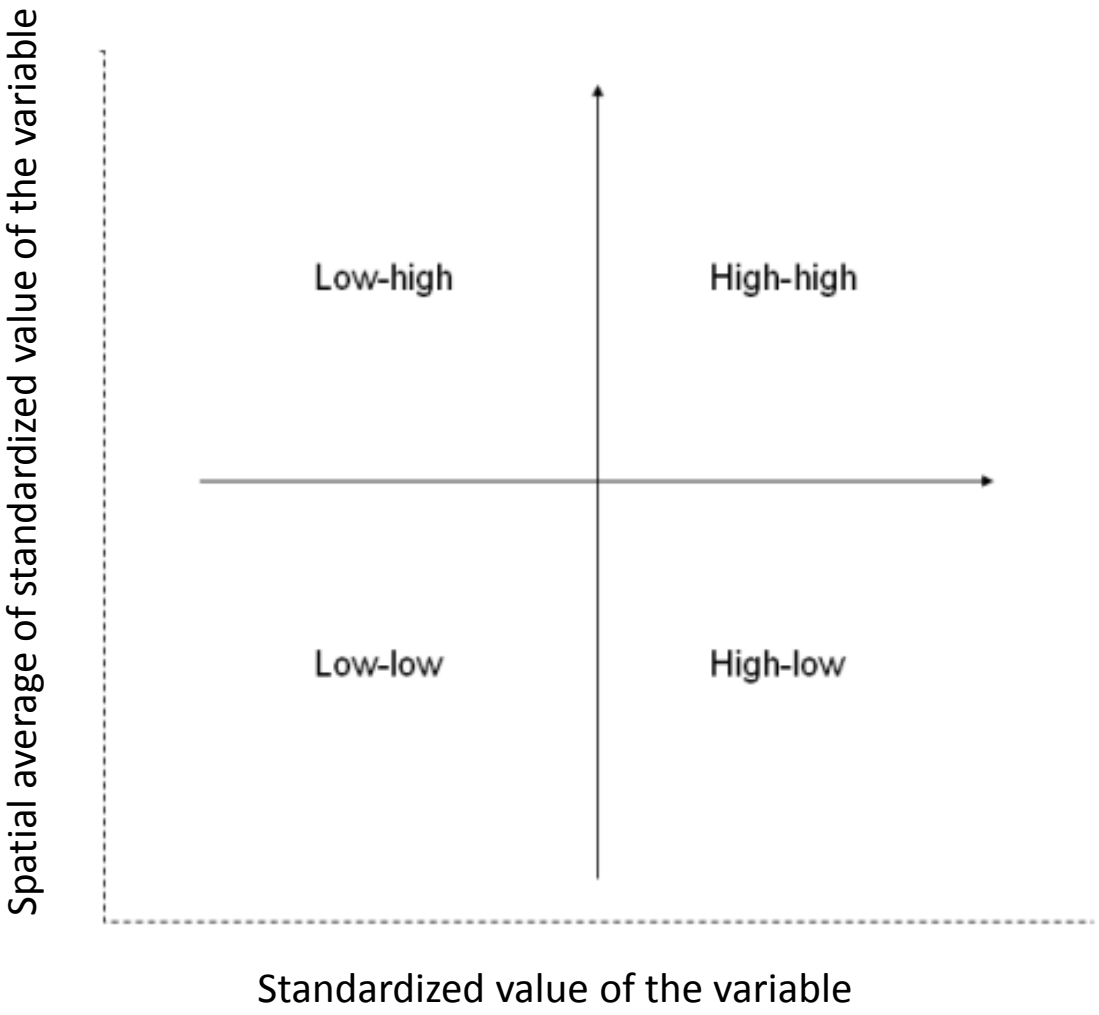
Moran scatter plot

Moran's I as a regression slope

Moran scatter plot: linear association between Wz on y-axis and z on the x-axis

Each point is pair (zi, Wzi), slope is I

Schematic presentation of Moran scatter plot



Exercise 8 (Moran scatter plot)

Same as exercise 7

Check the influence of outliers

- ± two standard-deviations

Use GeoDa menu:

Space > Univariate Moran's I

Data: Brazil_UF (Y00)

Inference

Null hypothesis: spatial randomness

- Observed spatial pattern of values is equally likely as any other spatial pattern
- Values at one location do no depend on values at other(neighboring) locations
- Under spatial randomness, the location of values may be altered without affecting the information content of the data

Computational inference

Permutation Approach

- Reshuffle observations – each observation equally likely to fall on each location
- Construct reference distribution from random permutations
- Normal approximation: $E(I) = -\frac{1}{n-1}$
- Not zero, but approaches zero as $n \to \infty$
- Pseudo significance

Exercise 9 (inference)

Illustrate the concept of spatial autocorrelation by contrasting real data that is highly correlated in space (% African-American 2000 Census tract data for Milwaukee MSA) with the same observations randomly distributed over space

Box maps and Moran scatterplots

Data: MSA (PCTBLCK, PCTBLACK)

Global *versus* local analysis

Global analysis

- One statistic to summarize pattern
- Clustering
- Homogeneity

Local analysis

- Location-specific statistics
- Clusters
- Heterogeneity

LISA definition

Local **I**ndicator of **S**patial **A**utocorrelation

Anselin (1995)

- Local Spatial Statistic
- Indicate significant spatial autocorrelation for each location

Local-global relation

- Sum of LISA proportional to a corresponding global indicator of spatial autocorrelation

Local Moran statistic

$$l_i = (z_i / m_2) \sum_j w_{ij}\, z_j$$

$$m_2 = \sum_i z_i^2, \qquad \sum_i l_i = nl, \qquad l = \sum_i l_i / n$$

Link local-global

Global is mean of locals

Inference

Computational

Conditional permutation

- Hold value at *i* fixed, permute others

LISA significance map

Locations with significant local statistics

Sensitivity analysis to p-value

Choropleth Map

Shading by significance

Non-significant locations not highlighted

LISA cluster map

Only the significant locations

- Matches significance map

Types of spatial autocorrelation

- Spatial clusters
  - high-high (red), low-low (blue)
- Spatial outliers
  - high-low (light red), low-high (light blue)

Spatial clusters and spatial outliers

Spatial outliers

- Individual locations

Spatial clusters

- Core of the cluster in LISA map
- Cluster itself also includes neighbors
- Use $p < 0.001$ to identify meaningful cluster cores and their neighbors

Caveats

LISA clusters and hot spots

- Suggest interesting locations
- Suggest significant spatial structure
- Do not explain

Need to account for multivariate relations

- Univariate spatial autocorrelation due to other covariates
- Spatial Econometrics

Exercise 10 (LISA)

Compute the local Moran statistic and associated significance map and cluster map

Assess the sensitivity of the cluster map to the number of permutations and significance level

Interpret the notion of spatial cluster and spatial outlier

Data: Brazil_MR (RENDAPC), MSA (PCTBLCK, PCTBLACK)

Reference

The notes for this lecture were adapted from those elaborated by Prof. Sergio Rey for the course "Geographic Information Analysis – GPH 483/598", held in the Spring of 2014 at the School of Geographical Sciences and Urban Planning at the Arizona State University.

They also relied on previous material prepared by Prof. Eduardo Haddad for the course "Regional and Urban Economics", held yearly at the Department of Economics at the University of Sao Paulo.